Exploiting sensor similarity to enhance data collection in massive IoT networks

Gwen Maudet

COLLEGES SCIENCES
BRETAGNE POUR L'INGENIEUR
LOIRE ET LE NUMERIQUE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPÉRIEURE
MINES-TÉLÉCOM ATLANTIQUE BRETAGNE
PAYS DE LA LOIRE – IMT ATLANTIQUE

ÉCOLE DOCTORALE 648
Sciences pour l'Ingénieur et le Numérique
Spécialité : *Informatique*

Par
Gwen MAUDET

Exploiting Sensor Similarity to Enhance Data Collection in Massive IoT Networks

Thèse présentée et soutenue à IMT Atlantique campus Rennes, le 23 novembre 2023
Unité de recherche : IRISA - UMR 6074
Thèse : 2023IMTA0360

Rapporteurs avant soutenance :

Alexandre GUITTON Professeur, Université Clermont Auvergne
Julien MONTAVONT Maitre de conférence, Université de Strasbourg

Composition du Jury :

Président :	Alexandre GUITTON	Professeur, Université Clermont Auvergne
Examinateurs :	Kinda KHAWAM	Maitre de conférence, Université de Versailles
	Julien MONTAVONT	Maitre de conférence, Université de Strasbourg
Dir. de thèse :	Laurent TOUTAIN	Professeur, IMT Atlantique
Co-dir. de thèse :	Mireille BATTON-HUBERT	Professeure, EMSE
Encadr. de thèse :	Patrick MAILLE	Professeur, IMT Atlantique

Invité(s) :

Yassine HADJADJ-AOUL Professeur, Université de Rennes

Abstracts

Abstract

The Internet of Things (IoT) is commonly employed for monitoring various physical quantities. In the innovative approach of Massive IoT (MIoT), a massive deployment of highly constrained sensors is considered to reduce deployment and maintenance costs. Aligned with this scenario, this thesis focuses on the development of mechanisms to reduce sensor energy consumption. The method relies on the principle of similarity: sensors can be considered similar if they provide similar observations. This approach enables the transmission of a subset of sensors to fulfill the monitoring requirements.

First, we identified and synthesized existing methods from the literature based on the principle of similarity. We established that this approach can be decomposed into three components, which we studied in the context of MIoT.

Next, we examined methods for managing sensor observations to maintain a constant stream of messages over time. Our first method involves having a specified number of sensors transmit in a round-robin fashion. The second method achieves precision results comparable to the first while reducing the number of sensor updates when the sensor fleet changes.

Finally, we propose a solution to form groups of sensors identified as similar by analyzing their observations. To this end, we introduce a new similarity measure based on interpolation, coupled with a hierarchical clustering method.

Résumé

L'Internet des objets (IoT) est couramment utilisé pour surveiller diverses grandeurs physiques. Dans l'approche innovante du Massive IoT (MIoT), un déploiement massif de capteurs très contraints est envisagé, afin de réduire les coûts de déploiement et de maintenance. Conformément à ce scénario, cette thèse se concentre sur le développement de mécanismes visant à réduire la consommation d'énergie des capteurs. La méthode repose sur le principe de similarité : les capteurs peuvent être considérés comme similaires s'ils fournissent des observations semblables. Cette approche permet la transmission d'un sous-ensemble de capteurs répondant aux exigences de surveillance.

Tout d'abord, nous avons identifié et synthétisé les méthodes existantes provenant de la littérature basées sur le principe de similarité. Nous avons établi que ce type d'approches peut être décomposé en trois composantes, que nous avons étudiées dans le contexte du MIoT.

Ensuite, nous avons examiné les méthodes de gestion des observations des capteurs permettant de maintenir une quantité constante de messages au fil du temps. Notre première méthode permet q'un un nombre spécifié de capteurs transmette en round-robin. La deuxième méthode atteint des résultats de précision comparables à la première tout en réduisant le nombre de mises à jour des capteurs lorsque la flotte de capteurs change.

Enfin, nous proposons une solution pour former des groupes de capteurs identifiés comme similaires en analysant leurs observations. À cet effet, nous introduisons une nouvelle mesure de similarité basée sur l'interpolation, associée à une méthode de regroupement hiérarchique.

Acknowledgement

Expressing sincere gratitude can be a challenge (possibly due to a sense of shyness?), but I will make a concerted effort to give my appreciation for all those who accompanied me on this journey.

I commence with my family and my parents, Martine and Eric, who have consistently offered their support and encouragement throughout the various stages of my personal and professional life. I have been fortunate to remain in close proximity to home (up to now...) which has allowed me to return and rejuvenate as needed, spending valuable time together, this closeness has been immensely beneficial. I also extend my thanks to my elder brother Tristan and my younger brother Erwan for the meaningful moments we've shared and those to come! I would also like to mention my two grandmothers, who I believe would be proud of their little one.

My heartfelt thanks extend to my circle of friends. The path of a thesis is a long and demanding journey, and the importance of taking breaks and recharge is crucial to avoid exhaustion. Our shared experiences have played a pivotal role in maintaining my equilibrium, and those involved will surely recognize themselves. From celebratory moments to convivial gatherings, adventurous escapades to sportive times.

Furthermore, I am appreciative of those who guided me towards this thesis. Alexander Pelov stands out for believing in me and offering a thesis position after my M1 internship at Acklio. The insights and feedback from my professors at IMT Atlantique regarding the thesis experience have been invaluable, as well as the advice from those who've tread this path before me.

A special thank you goes to the Chaire ValaDoE (VALeur Ajoutée DOnnées et Energie) and its affiliated partners for their financial support of this thesis project. Your sponsorship has allowed me to present my work at conferences without financial constraints.

My deepest gratitude is reserved for my mentor, Laurent Toutain, the director of this thesis. His ideas, vision, projects, and the freedom he provided have been instrumental in shaping a topic that resonated deeply with me. To my advisor, Patrick Maillé, your unwavering guidance, both in terms of formal proof methods and the enhancement of

my writing skills, have been invaluable. You have been an exceptional mentor. I also extend my thanks to Mireille Batton-Hubert, co-director of the thesis, for her enlightening discussions, attentive listening, and insightful perspectives on my progress.

My appreciation extends to Alexandre Guitton and Julien Montavont, the thesis reviewers, for undertaking the substantial task of evaluating three years' worth of research outcomes. My gratitude further extends to the entire jury, including Kinda Khawam and Yassine Hadjadjaoul, who provided assistance during the CSI sessions throughout these three years.

A special note of thanks to IMT Atlantique, for both my journey as an engineer (in Nantes) and as a doctor (in Rennes). I take pride in being a product of this merger. My gratitude also goes to the SRCD department, a warm and welcoming family, and to the numerous gatherings and interactions, formal and informal.

To my fellow ADERians, I am grateful for the camaraderie we've shared – whether in sports, over pizzas, during ice skating, or engaging in foosball or ping-pong matches. A special mention to my office mates: Ivan Marino ("BOFOU SAFOU"), followed by Ahn, Arnol Lemogué, my cherished companion and fellow traveler in conferences from Budapest to Cargese, and Pierre-Marie, a valued colleague and friend throughout this thesis journey (and beyond!). To fellow PhD candidates who started this journey alongside me: Amaury Bruniaux, Leo Lavaur, and Awaleh Houssein, I wish you all the best in your future endeavors!

To everyone, even those not specifically mentioned, I am genuinely grateful to have crossed paths during this significant phase of my life.

Table of Contents

Absctracts iii

Acknowledgement v

Table of Contents vii

INTRODUCTION 1

1 Introduction 3

1 Evolution in IoT . . . 3
2 A Massive IoT Deployment . . . 3
3 Problem Investigated in this Thesis . . . 4
4 Surveys on Sensor Efficiency and Observation Collection Management . . . 5
5 A Novel Approach to Managing Observation Collection Based on Similarity 6
6 Outline . . . 7
7 List of Publications . . . 8

I FRAMEWORK FOR SIMILARITY-BASED SENSORS EFFICIENCY 11

2 Specifications of the MIoT Scenario 13

1 Environment, Deployment Strategy, Objective . . . 13
2 Communications and Protocol . . . 15
3 Observations and Similarities . . . 17
4 Definition and Purposes of the Observation Collection Scheme . . . 18
5 Sum-up of the Points of Interest for an Observation Collection Scheme . . 19

3 A Survey on Data Collection Based on Sensors Similarity 20

1 Overview of the Papers Examined in the Survey . . . 20

1.1 Similarity Metric Definition 20
1.2 Data-driven Sensor Observations Scheduling 21
1.3 Coverage Problem 22
1.4 Efficient Placement 23
1.5 Compression of Transmissions 23
1.6 Fault Detection 24
2 Three Components to Define an Observation Collection Scheme 24
3 Similarity Metric 28
3.1 Similarity Metrics Based on Geographical Proximity 28
3.1.1 Distance Estimation Using Network Communication . . . 29
3.1.2 Inter-sensor Distance 29
3.1.3 Similarity Based on Sensing Range Modeling 29
3.1.4 Random Processes for Spatial Variations Modeling 31
3.2 Similarity Metrics Based on Observation Histories 32
3.2.1 Representing the Observation History as a Set of Values . 33
3.2.2 Representing the Observation History as an Ordered Vector 34
3.3 Discussion on the Choice of the Similarity Metric 35
4 Covering Subset Algorithm 39
4.1 Search for a Covering Sensor Subset for Coverage Problem 41
4.2 Partitioning into Covering Subsets 41
4.2.1 Partitioning the Sensor Field for the Coverage Problem . . 41
4.2.2 Graph-based Partitioning for Sensor Observations Scheduling 42
4.3 Clustering Similar Sensors 43
4.3.1 Graph-based Clustering 44
4.3.2 Clustering for Fault Detection 44
4.3.3 Clustering for Sensor Placement 44
4.4 Discussion on the Choice of the Covering Subset Algorithm 45
5 Activation Allocation Method 48
5.1 One Covering Set in Active Mode 49
5.1.1 Same Active Subset Until a Sensor Fails 49
5.1.2 Regular Updates of the Active Sensor Set 50
5.2 Round-Robin 51
5.2.1 Round-robin between Disjoint Covering Subsets 51

5.2.2 Round-robin within Each Cluster of Similar Sensors 51
5.2.3 Translation of a Round-Robin Method into Activation Period Updates 52
5.3 Discussion on the Choice of the Activation Allocation Method . . . 53
6 Conclusion 55

II ACTIVATION ALLOCATION METHOD 57

4 Selecting a Subset of Sensors in Round-Robin 59
1 Problem Statement and Model 59
1.1 Assumptions and Notations on Sensors 59
1.2 Formalization of the Period Allocation Function 60
1.3 Ensuring Regular Observations from a Sufficient Number of Sensors 61
1.4 Definition of the Quality Metric 61
1.5 Definition of the Monitoring Duration 62
2 Ensuring Periodic Activations From at Most M Sensors 64
2.1 Definition of Effectiveness 64
2.2 Overall Principle of the Period Allocation Function 65
2.3 Formal Definition of the Period Allocation Function 67
2.4 Properties of $f_{M,\tau}$ 70
3 Simulations 71
3.1 Influence of the Number M of Sensors Jointly Activating 71
3.2 "Diversity Versus Duration" Trade-Offs 72
4 Conclusion 74

5 Asynchronous 2-Level Round-Robin Activations 75
1 Problem Statement and Model 76
2 The Synchronized Round-robin Allocation Function 77
2.1 Definition of the Round-robin Function 77
2.2 Properties 78
3 The 2-Level Round-Robin Allocation Function 79
3.1 Functioning Principle of the Allocation Method 79
3.2 Construction of the Period Allocation Function 80
3.3 Properties 81

4 A Markovian Model for Performance Evaluation 82
4.1 Modeling Sensor Arrivals and Departures 82
4.2 Performance Metrics Estimation For 2LRR 83
4.2.1 Number of Sensors in the Steady State 83
4.2.2 Number of Period Changes 84
4.2.3 Mean Diversity 84
5 Simulation Results 84
5.1 Comparative Performance Evaluation 85
5.1.1 Simulation Setting 85
5.1.2 Other Scheduling Methods for Comparison 85
5.1.3 Performance Evaluation 86
5.2 Search for the Optimal τ Parameter 88
6 Conclusion and Discussions 89

III SIMILARITY METRIC AND COVERING SUBSET ALGORITHM 91

6 Grouping Sensors Based on Observations 93
1 Sensors, Observation Histories, and Objectives 93
1.1 Identifying Sensors Belonging to the Same Phenomenon 94
1.2 Incoming and Outgoing Sensors 94
1.3 Observations Sent by the Sensors 94
1.4 Observation History Definition 95
2 Defining a Distance Metric Based on Sensor Observations 95
2.1 Interpolation Function Based on an Observation History 96
2.2 Distance Based on Mean Magnitude Difference 99
3 Tuned Linkage Hierarchical Clustering 100
3.1 Specification of the Clustering Problem 100
3.2 The Agglomerative Hierarchical Clustering 101
3.3 Tuning of the Linkage Method 102
4 Simulations 103
4.1 Generation of Phenomena and Observation Histories 103
4.2 Kriging Parameter Settings 105
4.3 Evaluation of the Similarity Metric 106

4.3.1 Comparative Distance: Jaccard Distance 106
4.3.2 Performance Evaluation 107
4.4 Evaluation of the Clustering Solution 109
4.4.1 Comparative Solution: Limiting the Maximum Distance Within a Cluster 110
4.4.2 Threshold of 5 Clusters for the Agglomerative Hierarchical Clustering 111
4.4.3 Performance Evaluation 112
5 Conclusion and Perspectives 115

CONCLUSION 117

7 Conclusion and Research Directions 119
1 Summary of our Works 119
2 Limitations and Perspectives 122
2.1 Need to Address All the Fixed MIoT's Points of Interest 122
2.2 Benefits of Establishing a Testbed 122
2.3 Generalizing the Problem: Managing Heterogeneous Sensors 123
2.3.1 Different Networks, Different Data Formats 123
2.3.2 Integration of Sensors Connected to an Energy Source . . 124
2.3.3 Diverse Transmission Efficiencies 124
2.3.4 Varying Sensor Accuracy 124

APPENDIX 127

A Résumé en Francais 129

B Proofs of Bounds and Effectiveness in Chapter 4 133
B.1 Upper Bound of an Effective Period Allocation function 133
B.1.1 Preliminaries 133
B.1.2 Demonstration of Eq. (4.2) (Page 65) 134
B.2 Effectiveness of $f_{M,\tau}$ Over the Instants of Period τ 135
B.2.1 Sensor Representation Set 135
B.2.2 Characterization of the Sensor Set Over the Instants of Period τ . . 137

B.2.3 Preliminaries and Proof Scheme for the Effectiveness of $f_{M,\tau}$ 138
B.2.4 Demonstration of Proposition 2 (Page 70) 139
B.3 Bounds of the Sample Span of $f_{M,\tau}$ 141
B.3.1 Preliminaries to the Proof of the Lower Bound 141
B.3.2 Demonstration of the Lower Bound of $f_{M,\tau}$ 142
B.3.3 Demonstration of the Upper Bound of $f_{M,\tau}$ 145

C Proofs of Diversity Properties of Synchronous Round-Robin in Chapter 5 146
C.1 Same Activation Periods to Maximize the Mean Diversity 146
C.2 Regular Time Stamps to Maximize Minimum Diversity 148

D Calculations for Simple Kriging Resolution 152

BIBLIOGRAPHY 153

List of Figures

1.1 Illustration of an apartment with integrated IoT sensors: each red dot represents a sensor. 4

1.2 Taxonomy of IoT efficient mechanisms, from [RBC14] 6

2.1 Environment, sensors, communications 14

2.2 Modes of a sensor and modification of its activation period 16

3.1 Similarity quantification between one sensor and other sensors according to three levels of similarity: low, medium, high. 28

3.2 Inter-sensor distance [Bah+14], sensing range, thresholded similarity [KS17], common area metric [SSV12; Mam14; SP01; CW06] and sponsored sector [TG02; TG03] illustrations 30

3.3 Generic representation of an observation history. The observation history θ is represented as orange diamonds. The horizontal axis represents time, and the vertical lines indicate the borders for the sliding window until the last observation. 33

3.4 Sensors similarity graph based on the illustration shown in Fig. 3.1. A link is established between two sensors if the similarity metric exceeds the high similarity threshold (illustrated in green in the previous figure). 40

3.5 Disjoint Dominating Set generated from the sensor similarity graph in Fig. 3.4. Three disjoint subsets of sensors are displayed, and each subset is dominant. 42

3.6 Similar sensor Clusters based on the illustration shown in Fig. 3.1. The method depicted in the figure groups sensors together based on a high level of similarity, also relaxing by allowing the grouping of sensors with medium similarity. 43

3.7 Activation allocation method where always the same active sensor subset is chosen. The figure is based on the graph-based DDS resolution shown in Fig. 3.5, with the active subset corresponding to one of the covering sets. The sensors in the chosen covering subset activate at regular intervals, while all other subsets of sensors remain in deep sleep mode. 49

3.8 Three sensors transmitting in a round-robin manner: each sensor's observation is represented by colored squares. The result shows the set of observations made by the three sensors: receptions of observations at regular intervals, evenly distributed among the three sensors. 52

4.1 Evolution of freshness over time for 2 sensors in (a) and the corresponding diversity in (b). The x-axis represents time, with the observations sent by the sensors labeled as $S1$ and $S2$. The y-axis represents the freshness (or the sum of freshness) according to the observations, with the exponential freshness function. 63

4.2 Illustration of sensor activations using the function $f_{M,\tau}$ with parameters $M = 3$ and $\tau = 1$. In this scenario, we have a set of 7 sensors, each with equal battery capacities ($e = 10$) and activation and period change consumption ($c_e = c_r = 1$). These sensors become active at random times between $t = 0$ and $t = 25$.
. 66

4.3 Representation of some performance indicators using the period allocation function $f_{M,\tau}$, varying the number of jointly activating sensors M, for a few target activation periods τ. (a) corresponds to the sample span with analytical bounds, (b) the diversity. Each curve show values obtained over 100 simulation runs. 72

4.4 Performance of the update period function $f_{M,\tau}$, for several values of M (given in the legend) and τ (from 0.5 to 10 by 0.1 increments). Each point corresponds to the two-dimensional performance metrics (Diversity on the x-axis and monitoring duration on the y-axis), for fixed parameters M, τ of the update function. 73

5.1 Evolution of the binary tree representation as sensors enter *(top)* or leave *(bottom)* the system; each sensor is represented with a colored circle with an ID, and horizontal dotted lines represent the activation periods of the sensors at that depth. A dotted line around a sensor means that its position (and height) was changed in the tree (hence a period change order is needed).
The top part represents the successive arrivals of sensors indexed from 1 to 5; the bottom one shows the successive departure of sensors 4, 2, and 3 (departures are symbolized by a cross). 80
5.2 Continuous Markov modeling of the number of active sensors over time . . 83
5.3 Diversity over time for a simulation, with global period $\tau = 0.1$s for round-robin methods and individual period $p = 150$s for the static method, and diversity guarantee (fifth percentile) for each method (horizontal lines). . . 86
5.4 Fifth percentile of diversity (a) and number of period change per time unit (b) for the three methods, versus their parametrization. Parameters for round-robin methods are in bottom x-axis and in top x-axis for static. . . 87
5.5 Simulation versus theoretical results, with constant parameters for sensor arrivals and departures. One simulation trajectory of diversity over time is shown in (a), with the corresponding steady-state mean value for the Markovian model, for $\tau = 5$. In (b) we compare the fifth percentile of simulated diversity with the theoretical mean diversity: both reach their maximum for approximately the same global period τ. 89

6.1 Representation of two observations histories. 95
6.2 Diagram illustrating the variogram model based on experimental variogram points. The variogram consists of three parameters: nugget, sill, and range. 97
6.3 Orange diamonds and dashed green squares represent two observation histories, with time on the x-axis and observation values on the y-axis. The interpolations are depicted as solid orange and dashed green lines, respectively. The vertical dashed lines indicate the common temporal domain of the two interpolations $[a(.,.), b(.,.)]$. The distance between the two interpolations is defined by the mean distances (red arrows) over the common definition interval. 99
6.4 Examples of linkage methods, from [Gue11]. 101
6.5 Phenomena: (a) in their entirety, (b) zoomed between $t = 0$ and 200. . . . 104

6.6 Median, 10^{th} and 90^{th} percentiles of distances between sensors either from the same or different phenomena.
(a): Using our metric based on interpolation and mean amplitude difference.
(b): Using the Jaccard distance. 108
6.7 Similarity between ground truth and a clustering based on the observations histories. Linear regression of performance for $\sigma < 0.3$. 114

B.1 Representation of the number of period changes depending on the index of the sensor when $M < \frac{n}{2}$ 143
B.2 Representation of the number of period changements depending on the index of the sensor when $M \geq \frac{n}{2}$ 144

C.1 Diversity over time when 5 sensors have the same activation period and are scheduled to receive observations at regular intervals. Time is represented on the x-axis, with points S0, S1, S2, S3, and S4 indicating sensor activations; diversity is shown on the y-axis. 149
C.2 Diversity when sensors have the same activation period, moving $S3$ to the right. 150
C.3 Diversity among sensors with the same activation period, moving $S3$ to the left. 150

INTRODUCTION

Chapter 1

Introduction

1 Evolution in IoT

An IoT (Internet of Things) network comprises sensors deployed in an environment to monitor various physical quantities, such as temperature, humidity, or CO2 levels [Aky+02]. Its applications span across multiple domains, including healthcare, agriculture, environment, public safety, military systems, transportation systems, and industry [RBC14; GTG22]. The goal of an IoT network is to gain insights into an initially unknown environment in order to control actuators, or identify anomalies, for instance.

Currently, there are approximately 15 billion IoT devices[1], and it is projected that the number of connected objects will reach 500 billions by 2050[2]. To support the large-scale deployment of these sensors, energy autonomy is a desired characteristic. Additionally, it is crucial to maximize the lifespan of the network. As a result, significant research efforts have focused on reducing the energy consumption of sensors during communication [Per+20b; JC19; Per+20a; RC23].

2 A Massive IoT Deployment

Traditionally, a limited number of highly reliable sensors are strategically placed to gather targeted information. However, this approach suffers from an overreliance on each individual sensor. Indeed, each sensor needs to be carefully positioned to provide relevant observations. Moreover, if a sensor fails, it must be immediately replaced.

Today, the production cost of a constrained sensor, i.e., one with limited battery, computational power, and memory, is as low as 1$ per unit. This affordability enables the deployment of a large quantity of such sensors, for instance in everyday objects. This scenario falls under the paradigm of Massive IoT [Stu+19; Jou+23]. For instance,

[1] https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/
[2] https://emarsonindia.com/wp-content/uploads/2020/02/Internet-of-Things.pdf

[Mot+20] propose a perspective on managing large-scale outdoor air quality by considering a multitude of nodes, ranging from objects carried by individuals to vehicles. Additionally, in the context of smart home, an illustration of a connected apartment, where a majority of objects are furnished with IoT sensors, is presented in Fig. 1.1.

Figure 1.1: Illustration of an apartment with integrated IoT sensors: each red dot represents a sensor.

By increasing the number of observers in the environment and reducing their individual importance, this approach contributes to reduce deployment and maintenance costs. While this proposal offers enhanced versatility and enables the further expansion of IoT solutions, it also introduces new challenges in managing a fleet of sensors efficiently [Jou+23].

3 Problem Investigated in this Thesis

Drawing inspiration from our connected apartment example, a non-slaved solution would involve the sensors measuring the physical quantity (such as temperature) at regular inter-

vals and directly transmitting it to the monitoring system. While this approach effectively manages the studied environment, it demands a substantial sensor energy consumption.

Some deployed sensors are positioned in such close proximity that they generate highly similar observations. These redundancies can be identified and exploited to minimize sensor transmissions, thereby conserving their energy.

This thesis delves into effective collection methods suitable for the random deployment of a substantial number of constrained sensors. The methodology we explore is grounded in the assessment of similarity among sensors, which enables the reduction of transmitted data. For instance, one potential approach could involve relying solely on a subset of sensors to meet the required monitoring objectives without significantly compromising tracking accuracy.

4 Surveys on Sensor Efficiency and Observation Collection Management

The issue of energy efficiency is at the core of IoT challenges, and numerous solutions have been proposed and categorized in various surveys [Kum+17; CH20; Zan+21; Azi+13; Cor+07; DBO17; Eng+18; KL05; KAT12; Fas+07; MM08; NC23; CS16; WX06; Zhu+12; YA08]. A taxonomy proposed in Fig. 1.2 from [RBC14] highlights various approaches in this regard.

In this thesis, we focus on methods that aim to reduce the number of messages transmitted by the sensors, as transmission represents the primary factor of energy consumption [Ali+09; Ana+09]. Some surveys compiled methods that add intelligence to the node, such as sensing reduction methods based on predictions [DBO17; Eng+18] or message compression [KL05; KAT12; Fas+07; MM08; NC23]. Other surveys investigate strategies for placing sensors to enable more efficient message transmission [CS16; WX06; Zhu+12; YA08].

However, the sensors under consideration possess constrained memory and computational capacities, which restricts their ability to accommodate intelligent algorithms. Furthermore, in the context of a MIoT solution, the main objective is to minimize the human costs associated with strategic deployment. As a result, these introduced energy-efficient methods are not applicable to our context.

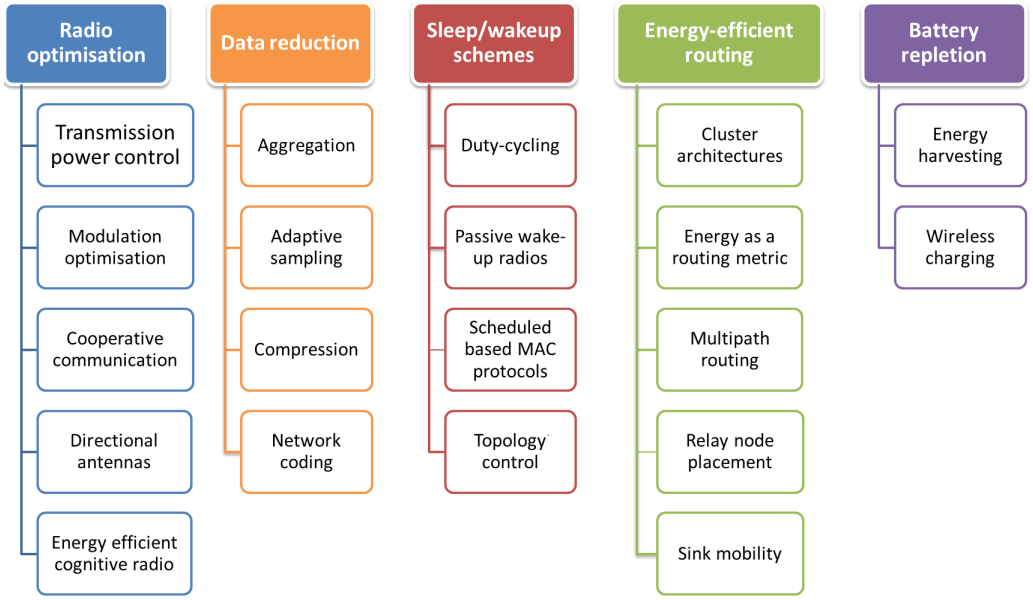


Figure 1.2: Taxonomy of IoT efficient mechanisms, from [RBC14]

5 A Novel Approach to Managing Observation Collection Based on Similarity

A fundamental assumption for MIoT scenarios is the deployment of a large number of nodes within the environment. Due to this high density of sensors, certain sensors are likely to produce observations that are closely aligned.

As a result, the establishment of similarity connections based on geographic proximity and/or the provision of analogous observations becomes feasible. Utilizing these similarity links, it becomes possible to reduce the required sensor set to be used for completing monitoring requirements, while the others can be in deep sleep to save their energy. This approach significantly curtails the volume of transmitted observations, resulting in substantial conservation of battery resources.

While this kind of approach has been proposed in some research papers, to our knowledge, no formalization of this concept has been provided. We propose a formalization of this methodology, structured into three key components. Specifically, these components encompass:

- A **similarity metric**, which is a positive real-valued measure that quantifies the proximity between sensors according to gathered information.

- A **covering subset algorithm**, leveraging the identified sensor similarities to generate one or more sensor subsets. Each subset fulfills the environmental monitoring requirements.

- An **activation allocation method**, which, based on the covering subset(s), distributes the observation load among the sensors.

In the upcoming part, we present a survey of solutions from the literature that are based on the similarity concept. Through our analysis, we show that all the solutions from the literature can be described using these three components.

6 Outline

The remainder of this thesis is split into three parts.

• Part I delves further into the problem under investigation in this thesis.

In Chapter 2, *Specifications of the MIoT Scenario*, we provide an in-depth exploration of our interpretation of a MIoT deployment. This encompasses considerations such as the types of sensors being considered, the network architecture envisioned, deployment specifics, and the surrounding environment. This leads to the key aspects that an observation collection management approach based on similarity should adhere to.

In Chapter 3, *A Survey on Data Collection Based on Sensors Similarity*, we present the first comprehensive survey focused on techniques for managing observation collections through the utilization of similarity principles. Specifically, we analyze the choices proposed in existing research for each of the three components, and evaluate the propositions put forth in the literature based on points of interest for the development of MIoT. Based on these findings, we propose novel approaches to overcome these limitations in the subsequent chapters.

• Part II investigate the period allocation component, where the focus is distributing the load of observations among a set of sensors, when sensors come and go.

Within Chapter 4, *Selecting a Subset of Sensors in Round-Robin*, we present a formalization of the activation allocation method, as the update of a sensor's transmission period after it has transmitted a message. Moreover, we introduce a method that facilitates a fair distribution of observation loads, ensuring the reception of observations at strictly defined time intervals (the first parameter of the function), distributed among

no more than a specified number of sensors transmitting in a round-robin manner (the second parameter of our function).

In Chapter 5, *Asynchronous 2-Level Round-Robin Activations*, we propose a second method that relaxes the strict message reception requirement and adapts to changes in the sensor fleet with minimal associated costs. We demonstrate that while this solution is theoretically sub-optimal, it closely approximates the performance of our initial solution. Moreover, it is better suited to the inherent constraints of the MIoT environment.

• Part III consists of a single chapter that delves into the two other fundamental components of the core method: the similarity metric and the covering subset algorithm.

In Chapter 6, *Grouping Sensors Based on Observations*, our focus is on the clustering of similar sensors based on a similarity metric derived from their observations. The similarity metric utilizes the Kriging interpolation method to compute the average magnitude difference between interpolations over shared time intervals. For the clustering process, we devise a hierarchical approach that takes into account the common period of presence of two sensors as a weighting factor in the linkage method. Through simulation, we demonstrate the superior performance of our proposed methods compared to state-of-the-art reference methods.

• Finally, we conclude this thesis in Chapter 7, where we propose future research directions.

7 List of Publications

International Conferences

* **Gwen MAUDET**, Mireille BATTON-HUBERT, Patrick MAILLE, Laurent TOUTAIN, *Emission Scheduling Strategies for Massive-IoT: Implementation and Performance Optimization*, IEEE/IFIP Network Operations and Management Symposium, 2022

* **Gwen MAUDET**, Mireille BATTON-HUBERT, Patrick MAILLE, Laurent TOUTAIN, *Energy Efficient Message Scheduling with Redundancy Control for Massive IoT Monitoring*, IEEE Wireless Communications and Networking, 2023

National Conferences

* **Gwen MAUDET**, Mireille BATTON-HUBERT, Patrick MAILLE, Laurent TOUTAIN, *Reduction de la redondance de messages des capteurs dans un contexte Massive IoT*,

25èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications, 2023

Part I

FRAMEWORK FOR SIMILARITY-BASED SENSORS EFFICIENCY

Chapter 2

Specifications of the MIoT Scenario

This chapter serves as an introduction to the fundamental concepts and terminology (highlighted in **bold**) that will be utilized throughout the thesis.

1 Environment, Deployment Strategy, Objective

Overall Objective

The objective is to monitor an **environment** (an agricultural field, a building for instance), by measuring a **physical quantity** (temperature, humidity, CO2). The ultimate goal can be to control this environment, possibly with actuators, although this thesis is only limited to the measurement task.

Traditional Approach VS MIoT Paradigm

In some monitoring applications, such as environmental monitoring experiments discussed in [US20], a **traditional approach** is adopted. This approach entails deploying on few strategic positions sensors characterized by high reliability, precise measurements, and large battery capacity. Implementing this approach requires the development of customized deployment strategies for each specific use case, often utilizing efficient sensor placement methods as presented in [CS16; WX06; Zhu+12; YA08]. Additionally, quick replacement of malfunctioning sensors is necessary. Generally, this approach heavily relies on each individual sensor, making it highly costly.

In this thesis, we adopt an alternative approach, based on the so-called **Massive IoT** paradigm [Stu+19; Jou+23]. This approach involves managing a large number of small, low-cost sensors with limited resources. For instance, in [Shi+01], a sensor density of up

to 20 nodes per cubic meter was envisaged. This approach overcomes the limitations of the traditional monitoring method: the deployment of MIoT is more convenient, as some sensors can be poorly placed without significantly impacting the system. Furthermore, there is no need for direct replacement of a battery-depleted sensor as long as there is another sensor remaining active nearby.

Deployed Hardware

A massive number of low-cost standard **sensors** are then deployed, each constrained by its energy, computational, and memory capacity, making them capable of performing only simple tasks [Mot+20]. These sensors can be distributed throughout the environment or integrated into objects. As time passes, new sensors may be added to the environment, while others may be removed or become inactive due to battery depletion.

At the other end, the **terminal** is the ending point for the transmitted messages. There is no energy limitation from the terminal's perspective. In Fig. 2.1, we provide an illustration of the sensors entering and exiting an environment and transmitting messages directly to the gateway (we will discuss the protocol later).

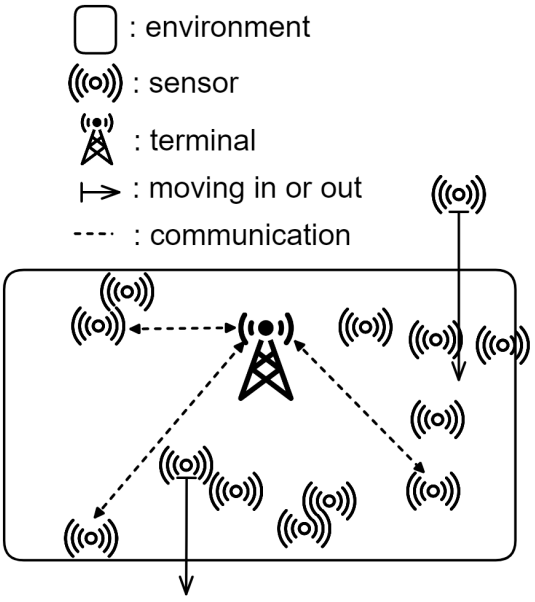


Figure 2.1: Environment, sensors, communications

Detailed Goal: Estimations for All Sensors at Target Observation Period

The term **estimation** refers to the evaluation of the physical quantity over time. For example, in a real-time temperature monitoring system, the system aims to receive observations at regular time intervals. For clarity, let's consider the **target observation period** as the desired interval for updating the estimations of all sensors.

Therefore, the objective is to generate estimations for sensors periodically, ensuring that each sensor's estimation is updated at regular intervals according to the target observation period. On the other hand, it is desired to extend the lifespan of sensors. This will be elaborated further, but the primary objective is to extend the lifespan of sensors in "densely deployed areas" considered more critical compared to zones where sensors are isolated.

2 Communications and Protocol

Communications Between the Sensors and the Terminal

Motivated by the emergence of LPWANs (Low Power Wide Area Networks) built to enable transmission over long distance with low consumption [Ban+22; Mek+19; BDK16], we assume that communication is possible only between the sensors and the terminal, and not among sensors [Kno06].

For this network, no message delivery guarantees are provided, which means that packets can potentially be lost in both sensor-to-terminal and terminal-to-sensor transmissions, as studied in [JR22; GZD21].

To curb any single node from monopolizing the radio spectrum, each object is constraint by a duty cycle, meaning it can only transmit for a fraction of the time (typically around 1% of the time) [Ade+17].

Sensor Modes and Protocol

A sensor operates in two modes, as illustrated in Fig. 2.2: the **active** mode, in which the sensor takes an **observation** of the physical quantity and transmits it, followed by a short listening period for period change orders from the terminal, and the **deep sleep** mode, during which it is inactive. The active mode is much more energy consuming than

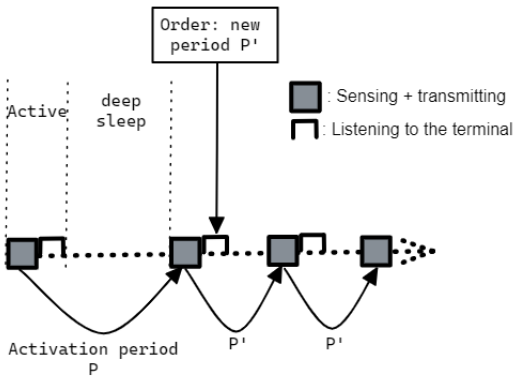


Figure 2.2: Modes of a sensor and modification of its activation period

the deep sleep mode: the energy consumption ratio between the two modes can be as high as 10^4 [Lal+18; Bou+18].

This represents the model of periodic wireless sensor networks as defined in [AA20]. The sensor is governed by a single parameter: the **activation period**, which corresponds to the duration between two active modes. The opening window after each sensor message is based on the LoRaWAN[1](Low Range Wide Area Network) Class A, an LPWAN created to be suitable for highly constrained sensors.

Each observation is directly sent to the terminal, insuring updates of fresh sensor estimations. Additionally, the periodic transmission mode and the reprogrammability of the activation period can be handled even for sensors with low algorithmic and memory capacity.

Limitations on Commands Issued by the Terminal

As previously discussed, the terminal is constrained by duty-cycle limitations, which allow only a small fraction of time for transmission. This constraint can become particularly restrictive when dealing with a considerable number of sensors that need to be managed.

During the time when the terminal is transmitting commands, it is unable to receive messages. Consequently, any observations transmitted by sensors within this timeframe are lost.

From the sensor's perspective, interpreting the need to adjust its activation period requires the sensor to enter an extended listening mode and employ signal processing methods, resulting in additional energy consumption.

[1] `https://lora-alliance.org/`

Due to these factors, it holds significance to effectively minimize the quantity of commands issued by the terminal.

3 Observations and Similarities

During its active mode, the sensor perceives a physical attribute from the environment, which it transmits to the terminal as an observation. This **observation** comprises attributes of both value and time.

Poor Quality Observations

The poor quality of the sensors makes the measuring tools imprecise, fragile and prone to failure [Kar+16]. This results in erroneous observation values in the form of drift, fixed error, or random error, for example [Elh+17; TC07].

Inexact Periodic Transmissions

The sensors transmit messages based on a periodic activation. However, due to their poor quality, malfunctions can occur on the clock, which can alter the exact duration of the deep sleep period [Tjo+04].

Physical Quantity Variations

The physical quantity exhibits spatial and temporal variations where certain typical behaviors may emerge, which we present in a non-exhaustive manner.

For instance, spatial variations can be very different from one location to another [Gru+06; Wan+13]. One can imagine an environment partitioned into separate spaces, such as the rooms of a building. Thus, the physical quantity may show small variations within the same room, but very large variations between two points that are close but in different rooms [LWP07].

Furthermore, spatial behavior can vary over time: two rooms may have a similar temperature on a cloudy day (winter) but significantly different temperatures on a sunny day (summer) [Gru+06; Mey+11].

Similarity Between Sensors

Under the assumption of dense deployment, some sensors may return similar kind of data, which has been observed in [Bur+09] in terms of the proximity of observation values, or in [Kai+16] in terms of geographical proximity. Therefore, the **similarity** metric is defined as the tendency for sensors to return the same observation values. In cases of strong similarity, sensors can be identified as **similar**: when a sensor transmits an observation, its estimation is updated, as can be the estimations of sensors that are similar to it.

4 Definition and Purposes of the Observation Collection Scheme

This thesis focuses on methods aimed at limiting the active mode of sensors. In the present network model, a so-called **observation collection scheme** is a method that can modify the activation period of a sensor that has just sent a message, if required.

Objective of the Observation Collection Scheme

In a system lacking an observation collection scheme, each sensor transmits messages at the target observation period. While this method fills the estimation requirements, it leads to a high volume of transmitted messages.

According to this concept of similar sensors, we can avoid activating a sensor if its estimation can be updated by a similar sensor. By leveraging this approach, we adapt the active mode of sensors by adjusting the activation period of a sensor after it has transmitted a message.

Leveraging Strong Similarities While Disregarding Isolated Sensors

The sensors are deployed throughout the environment, forming sets of sensors with high similarities, and others where sensors are only minimally or not at all similar to each other.

With the observation collection scheme presented here, we can significantly prolong the activation period of sensors in areas with high similarity among sensors, thus extending

their lifespan. On the other hand, a sensor with no similarity will continue to consume its energy at a high rate and will die quickly.

We consider that isolated sensors (with low similarity to other sensors) are poorly positioned, and therefore, our objective is to identify the sets of sensors with very high similarity to maximize their lifespan.

5 Sum-up of the Points of Interest for an Observation Collection Scheme

We can summarize the essential points of interest for the development of an observation collection scheme. These points will serve as guidelines to analyze the proposals from the literature, as discussed in Chapter 3, and will also guide us for the further proposed approaches in Chapters 4 to 6.

Regarding the study of the environment, we aim to investigate **complex environment** where the physical quantity varies differently across different zones. As a monitoring solution must endure over time, we strive to adapt to a **changing environment** where observed properties of the physical quantity fluctuate.

Our ambition lies in crafting a versatile solution capable of adapting to diverse scenarios. In terms of sensor deployment, we must grapple with **deployment issues**, which may include poorly placed sensors. Furthermore, we aim to accommodate **variations in the number of sensors**, encompassing the integration of new sensors as well as the potential exit or failure of sensors due to battery depletion.

We seek a solution that boasts resilience within constrained networks, where **packet losses** in both uplink and downlink are likely. When implementing the strategy, it remains important to **limit the number of commands issued by the terminal**.

Finally, we address the issue of poor sensor quality. This involves managing **clock drift**, characterizing the sensitivity of the measurement tool to deal with **corrupted sensors**, which may transmit aberrant observation values like random values, fixed errors or drift, and effectively handling **measurement errors**, which must not be neglected for these constrained objects.

Chapter 3

A Survey on Data Collection Based on Sensors Similarity

This chapter delves into energy-efficient solutions derived from existing literature, which harness the concept of similarity to efficiently distribute the observation workload among sensors.

Through a comprehensive investigation of papers in the literature that tackle this scenario, we demonstrate that a solution can be deconstructed into three core components. For each of these components, we conduct a comprehensive review of the existing literature. We analyze them from different perspectives crucial to a MIoT deployment, thereby shedding light on the limitations of previous works.

1 Overview of the Papers Examined in the Survey

Here, we introduce the papers that we will focus on in the following of the survey. These papers share a common scenario: a large number of sensors are deployed in an environment, allowing to develop a similarity metric to reduce the energy consumption of each sensor. For this, papers from diverse domains are relevant.

1.1 Similarity Metric Definition

Considering densely deployed sensors, some papers focus on studying the similarity between sensors. This similarity is defined in several manners.

In studying similarity, [SSV12] utilizes the concept of sensing range, which represents the radius within which a sensor can measure its surrounding environment. Consequently, the area covered by a sensor forms a circle (or sphere) centered around the sensor. Using this model, the authors construct graphs where the vertices are the sensors, and edges are established between sensors whose sensing areas intersect. They further investigate the

effects of varying the sensing range and highlight its influence on the constructed graphs.

[VAA04] develops a correlation model between sensors based on distance, employing statistical analysis. Gaussian process modeling is performed, assuming that the physical quantity is a random process. Particularly, the proposed model accounts for encoding performed by the sensor.

[APM09] explores the link between spatial similarity and similarity between observations of sensors. The goal is to validate spatial correlations by employing data-driven metrics.

1.2 Data-driven Sensor Observations Scheduling

Several methods have been devised to reduce the volume of transmitted messages through scheduling strategies among sensors, achieved by assessing the similarity based on the returned observations. These techniques establish similarity connections whenever two sensors transmit sufficiently similar observations. They further develop algorithms where, in each round, a subset of sensors is activated. This subset, through the expansion of their similarity connections, encompasses the entire array of sensors.

The ensuing research papers are the most closely related to the subject matter of this thesis.

In [CKJ05], a method is proposed to build groups of sensors that return highly similar observations, so that instead of all sensors being activated in each round, sensors from a same cluster are activated in a round-robin fashion. An experiment is conducted with light sensors placed under desk lamps and barriers positioned in certain areas. The method successfully groups together sensors surrounded by barriers, resulting in an average reduction of sensor consumption by a factor of 3 and low precision loss. The authors also present an extended simulation on a larger scale with similarly promising results. An extended version of this method is proposed in [LWP07], where sensors are randomly activating on time slots to create redundancy at the terminal end.

In [Liu+13b], similar principles are employed, with the transmission of information in a mesh network considered. To handle this last hypothesis, sensors near the terminal are more frequently activated to relay observations.

Another approach, presented in [KTP06], involves creating a transfer function that allows estimating the observation of one sensor based on the observation of another sensor. A link is established if the transfer function can accurately estimate the measurement of one sensor from its own measurement. An experiment is conducted with 54 temperature

and humidity sensors deployed with distances ranging from 6 to 15 feet. Through their experiment, up to 12 subsets of sensors are constructed, where each subset is able to update the estimation of the entire fleet of sensors, with the subsets transmitting in round-robin fashion. The results showed that this approach achieved a level of precision that is close to the scenario where all sensors are transmitting at each round.

1.3 Coverage Problem

Extensive research has been conducted on the coverage problem, and comprehensive explanations can be found in surveys such as [Meg+01; Zhu+12]. The objective is to achieve sufficient quality of service (called coverage) related to the spatial coverage of an environment, while minimizing the number of active sensors. The precise interpretation of the coverage depends on the sensor coverage modeling. This area has been widely studied, and we have herein focused on a subset of the proposed approaches.

The prevalent modeling considers sensors that cover the environment over a sensing range and the objective is to select a subset of sensors to optimally cover the environment (in totality, or with sufficient covering). [CW06] investigates a 3D environment, highlighting optimal conditions for sensor placement and developing a method for selecting a subset of sensors that ensures complete coverage. [TG03] presents a sensor selection method that is robust to packet loss, localization errors, and node failures through simulations. [AC16] addresses adaptation mechanisms in response to the death of a sensor by activating not used sensors. [Bah+14] also focuses on relay of depleted sensors. Specifically, they propose an efficient sleeping schedule for unused sensors that takes into account the fact that, as time progresses, a sensor is more likely to be selected in coverage task due to fewer available candidates.

[Mam14] assumes the same sensing modeling, and proposes a method that, at each scheduling step, choose one subset of sensor that insures coverage, in order to overall balance energy consumption across all sensors. [Mos+17] proposes a solution based on reinforcement learning, where a set of sensors is chosen at each step, while learning for future decision-making.

With the same range modeling for sensors, in [SP01], an algorithm is developed to construct multiple subsets of sensors, with each subset taking turns being active. Similarly, [Kra+11] suggests generating multiple covering subsets of sensors in conjunction with addressing the efficient placement problem, demonstrating that it is more effective to study both problems simultaneously rather than sequentially.

In [DB12], a statistical model based on geography is used for sensor coverage modeling. The solution allows selecting a minimum subset of sensors, subject to a minimum distortion constraint. Distortion is mathematically defined as the difference between the phenomenon and its estimation. [Raj11] also uses a statistical modeling, and presents an optimal sensor selection method based on an entropy metric.

1.4 Efficient Placement

The problem of efficient sensor placement aims to strategically position sensors in order to transmit observations efficiently and conserve energy. This topic involves establishing metrics to study similarities, and building algorithms for interpreting these similarities.

In the work presented in [Kra+06], the authors model the similarity between observations and the transmission cost using a Gaussian process over the spatial dimension. Utilizing this model, they propose a sensor placement heuristic that finds a balance between sending diverse observations and minimizing communication costs.

In [Yog+18; GGJ09], an initial dense sensor deployment is considered, and the objective is to reduce the number of sensors to be retained. [Yog+18] addresses this scenario by considering a large number of deployed sensors and aiming to select the most relevant ones. It is proposed an algorithm that clusters similar sensors based on both sensors observations and position; at the end, only one representative per cluster is retained in the solution. In an experiment, 30 sensors were deployed for relative humidity, luminance, and temperature measurements. It is demonstrated a reduction by four of the initially deployed number, with minimal loss of precision in the tracking of the physical quantities. In [GGJ09], the authors seek to decrease the number of sensors placed on the body for movement recognition. They propose a method where, for each sensor, a graph is built: the nodes represent different movements and an edge exists between two movements that can be distinguished by that sensor. They developed a method to find the smallest set of sensors such that each movement is completely distinguishable, equivalent to finding the minimum number of sensors for which the union of the graphs induced by this subset is fully connected.

1.5 Compression of Transmissions

The compression-based methods aim to reduce the amount of transmitted observations by sending a data representation model instead of the entire data set. The papers that

are linked to our topic propose to send a model composed of observations coming from multiple sensors.

In [TM06], the authors propose grouping sensors that provide similar observations and then sending an autoregressive model output that represents the combined observations of the sensor group. Similarly, [Alm+16] applies the same principle by reconstructing the linear evolution of the physical quantity for the considered cluster.

1.6 Fault Detection

The fault detection problem can be addressed by identifying similarities between sensors. In [Yoo20], correlation analysis is performed, allowing non-disjoint grouping of correlated sensors. Thus, a fault is identified when a sensor deviates from an assigned cluster. The authors conduct experiments on an industrial process monitored by 17 sensors and achieve better fault detection rates compared to some traditional methods.

2 Three Components to Define an Observation Collection Scheme

We have presented papers from various research domains. The two closely related research domains that extensively utilize similarity for reducing observation collection are *data-driven sensor observations scheduling* and the *coverage problem*. The other research domains focus on specific aspects of the problem or do not validate the main hypothesis: that observations made by sensors are directly transmitted to the terminal.

We propose to categorize these studied papers under a unified framework. This framework is centered around the core principle of leveraging similarity to curtail transmissions from sensors. The main goal is to transform sensor information (such as sensor position and sent observations) into a representation of similarity among sensors. Subsequently, based on these similarity relationships, an algorithm is used to construct one or multiple subsets of sensors, referred to as **covering subsets**, where each subset is built to fulfill the prescribed monitoring requirements. Finally, the distribution of observation load among sensors is defined.

This observation collection scheme can then be dissected into three principal components:

• Section 3 - **Similarity metric**: A real-value metric that quantifies the proximity between sensors, relying on known sensor information.
• Section 4 - **Covering subset algorithm**: Based on this similarity metric, one or more covering subsets of sensors are constructed, where each subset ensures the fulfillment of the requirements for monitoring the physical quantity.
• Section 5 - **Activation allocation method**: Building upon the covering subsets, this component defines how to distribute the load of observation transmissions among sensors.

We present the choices made for each paper regarding each of the three components in Table 3.1; for papers that address only a part of the problem, we indicate the missing components with ∅.

In the upcoming sections, we will delve deeper into each component by examining the specific approaches adopted by the reviewed papers. This exploration will be guided by identifying the most suitable strategies for IoT deployments.

Table 3.1: Studied contributions that reduce IoT sensor consumption based on similarity.
Presented following the three-component structure of an observation collection scheme.

Ref	Field	Similarity metric	Covering subset algorithm	Activation allocation method
[SSV12]	Similarity metric definition	Common area	∅	∅
[VAA04]	Similarity metric definition	Distortion	∅	∅
[APM09]	Similarity metric definition	Inter-sensor distance Jaccard coefficient Cosine similarity Pearson coefficient	∅	∅
[CKJ05]	Data-driven sensor observations scheduling	Max magnitude difference Similar trends counting	Partition into cliques	Round-robin
[LWP07]	Data-driven sensor observations scheduling	Max magnitude difference Similar trends counting	Partition into cliques	Round-robin
[Liu+13b]	Data-driven sensor observations scheduling	Max magnitude difference Similar trends counting	Partition into cliques	Round-robin
[KTP06]	Data-driven sensor observations scheduling	Isotonic regression function	Disjoint dominating sets	Round-robin
[CW06]	Coverage problem	Common area	One covering subset: heuristic	Same covering set in active mode
[TG03]	Coverage problem	Sponsored area	One covering subset: step-by-step	Regularly changing covering set in active mode

Table 3.1: Studied contributions that reduce IoT sensor consumption based on similarity.
Presented following the three-component structure of an observation collection scheme.

[AC16]	Coverage problem	Conflict parameter Spatial correlation	One covering subset: heuristic	Same covering set in active mode
[Bah+14]	Coverage problem	Inter-sensor distance	One covering subset: heuristic	Same covering set in active mode
[Mam14]	Coverage problem	Common area	One covering subset: heuristic	Regularly changing covering set in active mode
[Mos+17]	Coverage problem	Thresholded distance	One covering subset: step-by-step	Same covering set in active mode
[SP01]	Coverage problem	Common area	Partitioning into covering subsets: heuristic	Round-robin
[Kra+11]	Efficient placement Coverage problem	Mutual information	Partitioning into covering subsets: heuristic	Round-robin
[DB12]	Coverage problem	Distortion	One covering subset: heuristic	Same covering set in active mode
[Raj11]	Coverage problem	Entropy	One covering subset: heuristic	Same covering set in active mode
[Kra+06]	Efficient placement	Mutual information	∅	∅
[Yog+18]	Efficient placement	Mean magnitude difference Inter-sensor distance	K-means clustering Gaussian mixture model	∅
[GGJ09]	Efficient placement	Bhattacharyya distance	∅	∅
[TM06]	Compression	Max magnitude difference	Partition into cliques	∅
[Alm+16]	Compression	Max magnitude difference Similar trends counting	Partition into cliques	∅
[Yoo20]	Fault detection	Pearson correlation	Gaussian mixture model	∅

3 Similarity Metric

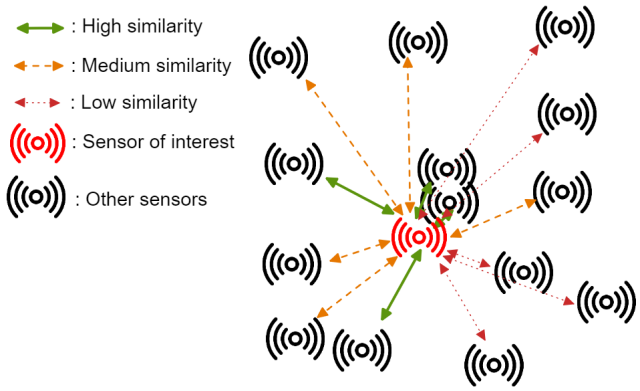


Figure 3.1: Similarity quantification between one sensor and other sensors according to three levels of similarity: low, medium, high.

This section describes similarity metrics that can be used to compare either individual sensors or sensor subsets. A similarity metric is defined as a positive real value, denoted by S, and the dissimilarity (or distance) metric is denoted by D, and is calculated from available sensor information. Fig. 3.1 illustrates the similarity between a sensor and the other sensors.

In the existing literature, we identify two main families of similarity metrics:

- Section 3.1 - **Based on the geography**: These metrics utilize the geographical positions of the sensors.
- Section 3.2 - **History data driven**: These metrics rely on the observations provided by the sensors.

We provide a comprehensive presentation of these families of metrics, reviewing the proposals from the literature. Hence, we compare them through points of interests for MIoT, with highlights of the open issues.

3.1 Similarity Metrics Based on Geographical Proximity

Similarity metrics based on geography consider that the spatial variations of the physical quantity are a function of the spatial dimension. In other words, sensors that are located closer to each other are more likely to send similar observations. This section presents various similarity metrics that are based on the geographic distance between sensors.

We denote the geographic distance between two sensors by d and d_i with a specific other sensor i.

3.1.1 Distance Estimation Using Network Communication

To use a similarity metric based on geography, it is supposed that the position of all sensors is known, equivalent to the distance between each pair of sensors. In a dense deployment of sensors, it is usually beneficial to use position estimation tools rather than registering the positions. Since the behavior of electromagnetic waves used for sensor communication can be mathematically modeled, it is feasible to utilize position estimation methods based on various network metrics [Gez08]. It is needed at least three terminals for which the position is known to evaluate the position of a sensor in 2D.

Trilateration can be used to determine the geographical coordinates of each sensor based on the estimated distances to the terminals. The Received Signal Strength Indication of a message can be modeled as an inverse-square law decay relatively to the distance, helping to estimate the sensor-terminal distance [OQ09]. The Time on Air can also be used to estimate this distance, possibly in conjunction with a clock signal or ultrasonic transmission [SHS01].

Other methods for estimating sensor positions rely on triangulation and exploit the Angle of Arrival [DKP19].

One of the most accurate, albeit resource-intensive, method for determining sensor positions is the GPS (Global Positioning System) and is based on the trilateration.

3.1.2 Inter-sensor Distance

One approach to measuring the similarity between sensors is to directly use their geographic distance. This dissimilarity metric, denoted as $D = d$, is employed between every pair of sensors, and is used in [Bah+14] for selecting a subset of sensors for the coverage problem.

3.1.3 Similarity Based on Sensing Range Modeling

Sensor similarity metrics based on geography are primarily used for the spatio-temporal *coverage problems*, as defined in Section 1.3. Given a sensor capable of covering the environment over a distance r (or r_i for sensor i) called the **sensing range**, the goal is to cover the entirety of the environment with a minimum number of sensors.

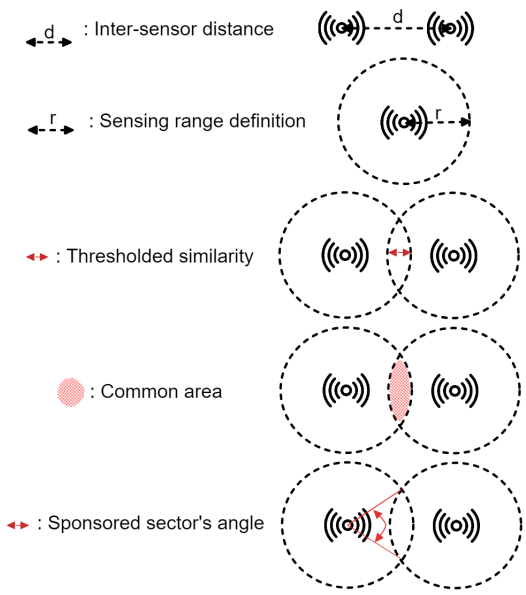


Figure 3.2: Inter-sensor distance [Bah+14], sensing range, thresholded similarity [KS17], common area metric [SSV12; Mam14; SP01; CW06] and sponsored sector [TG02; TG03] illustrations

Similarity Between Two sensors

To compute the similarity between pairs of sensors, [KS17] suggests establishing a threshold to the distance, constraining it to twice the sensing range: $S = \max(0, 2r - d)$; if two sensors are at least a distance of $2r$ apart, they have no similarity; the similarity is illustrated in Fig. 3.2. Alternatively, a binary version is proposed in [Mos+17], where similarity is assigned a value of 1 if the sensors are within a distance less than $2r$, and 0 otherwise.

Authors in [SSV12; Mam14; SP01] propose a method for calculating the exact proportion of common area between two sensors, as illustrated in Fig. 3.2. This similarity metric is defined as $S = \frac{\cos^{-1}(\frac{d}{2r})}{\pi} - \frac{d}{\pi(2r)^2} \times \sqrt{((2r)^2 - d^2)}$. Using union and intersection rules, it is possible to calculate the total area covered by a set of sensors, which is not possible with the method proposed in [KS17; Mos+17]. In [CW06], the problem is studied in three dimensions, where the observation of a sensor comprises a volume rather than an area.

Authors in [TG02; TG03] define a similarity metric that allows for different sensing ranges. They propose an asymmetric metric called the sponsored sector, which is defined as the airspace formed by the circular arc at the intersection of common areas (if they exist) and the segments directed toward the center of the sensor. This is illustrated in

Fig. 3.2. If we consider r_i and r_j as the coverage radius centered on nodes i and j, respectively, denoted as $S_{j\to i}$, the following conditions apply: • If $r_i + d < r_j$, then the sensing area of node i is entirely contained within the sensing area of node j, and the similarity is upper bounded to $S = 2\pi$. • If $d \leq r_j + r_i$, then the two sensing areas intersect, and the similarity is determined by the angle of the sponsored sector from node j to node i. Mathematically, this similarity is defined as $S_{j\to i} = 2\cos^{-1}\left(\frac{d^2+r_i^2-r_j^2}{2r_i d}\right)$.

Similarity Between One and a Set of Sensors

It can be relevant to evaluate a similarity metric for a sensor in relation to a set of sensors. In [AC16], the so-called conflict parameter is a dissimilarity metric which quantifies how much the sensor has common sensing range with other sensors: $D = \sum_{j\in\text{sensors}} \frac{\min(d_j,2r)}{2r}$. Moreover, a second metric called the spatial correlation parameter is presented, measuring the distance of the sensor from active sensors without bounds by the sensing area: $D = \sum_{j\in\text{active sensors}} \frac{d_j}{(\sum_{\text{active sensors}})^2}$. This latter metric allows differentiating between two sensors that have the same conflict parameter. A metric proposed by [Mam14] is the shared sensing region, calculated from the similarity matrix based on the common areas between pairs of sensors. This latter metric serves to define precisely the usefulness of a sensor in relation to the other active sensors.

3.1.4 Random Processes for Spatial Variations Modeling

The use of probabilistic models to translate the functioning of an environment allows obtaining exploitable theoretical results. Spatial variations of a physical quantity can be modeled through a Gaussian process described by a mean function and a covariance function. Variogram-based methods are then used to specify the correlation function [Cre93]. The correlation between sensors is a dissimilarity metric defined through their distance d and is denoted $\hat{\gamma}(d)$. For example, the exponential kernel is written $\hat{\gamma}_{\sigma,r}(d) = \sigma e^{(-d^2/r)}$, the Matérn 3/2 is written $\hat{\gamma}_{\sigma,r}(d) = \sigma(1+d)e^{-d/r}$ [Dur01], where σ is the variance parameter and r is the length-scale parameter.

Based on this modeling framework, various similarity metrics can be employed. For example, [Raj11] utilizes an entropy criterion, while [Kra+06; Kra+11] define the mutual information retained for a subset of sensors. [VAA04; DB12] employ a metric known as distortion, taking encoding into the statistical model.

3.2 Similarity Metrics Based on Observation Histories

By utilizing historical transmission data, it is possible to quantify the proximity of two sensors through the application of similarity metrics that are based on the sensors' observations.

An **observation** comprises a **value** θ_t captured at **time** t, such that an observation history is defined as $\theta = \{\theta_t, t \in T\}$, where T represents the set of times at which observations are recorded.

Sliding Window Definition

It is possible to compute the similarities between sensors at the firsts steps of the monitoring process, assuming that the computed relationships between sensors remain static throughout the monitoring duration. However, changes in the environment or behavior of individual sensors can impact the relationships between sensors. In order to account for these changes, continuous computation of similarities should be performed to dynamically update the observation collection scheme accordingly.

To determine the time frame in which the recorded observations are retained, a **sliding window** approach with a duration of Δt is generally employed [Tao+23; Chu95; Vaf+14; CN16; KGG11]. Specifically, considering the current time $\tilde{t}$, only the observations made after $\tilde{t} - \Delta t$ are considered in the analysis: $\{\theta_t, \tilde{t} - \Delta t \leq t \leq \tilde{t}\}$.

In general, selecting a larger value of Δt is advantageous in terms of improving similarity robustness. On the other hand, giving too much weight to outdated data can reduce responsiveness. If sensors with similar behavior begin to return vastly different readings, a smaller value of Δt is preferred to accurately characterize this deviation. Thus, a judicious choice of Δt must be made to balance robustness and responsiveness to changing environments, as can be discussed in [Che+12; KPK97; Chu+15] in related topics.

Specificities of the Observation Histories

Due to the continuous computation of similarity, the observation collection scheme adjusts the transmission period of sensors over time. This, coupled with possible packet losses, results in an irregularly sampled observation history between sensors.

Moreover, achieving a scenario where all sensors activate only on defined time stamps is in general not feasible. Sensors are susceptible to clock drift, which slightly alters their real activation periods, rendering the synchronization of a sensor fleet complex.

Additionally, the use of low-cost sensors introduces noise during the sensing process, which can make similarity calculation unreliable [Kar+16].

In summary, sensors transmit messages in an irregular manner, synchronization of transmissions on the same time slots cannot be guaranteed, and sensor observations are affected by noise. A general scheme of an observation history is presented in Fig. 3.3.

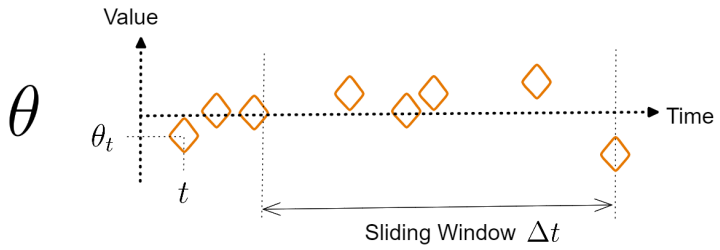


Figure 3.3: Generic representation of an observation history. The observation history θ is represented as orange diamonds. The horizontal axis represents time, and the vertical lines indicate the borders for the sliding window until the last observation.

Summary of History Data Driven Metrics

Different metrics between observation histories are proposed in the literature. In particular, we have classified them according to the way the observation histories are represented:

- Section 3.2.1 - **Representing the observation history as a set of values**: erasing the temporal dimension of the observations to represent the observation history only through the observation's values, in order to use statistical sample metrics [PT00].
- Section 3.2.2 - **Representing the observation history as an ordered vector**: assuming that the i^{th} observation's time of all observation history are taken at the same time, to use vector-based metrics [AC17; NWN15].

We do not specify whether the similarity metric is taken between one sensor and another, or from one sensor to a set of sensors. The similarity metrics are based on observation histories, which can be derived from one or multiple sensors.

3.2.1 Representing the Observation History as a Set of Values

In this part, the similarity between observation histories θ_i and θ_j is made through the comparison of their two sets of values. Through this modeling approach, observation

histories do not require any assumptions about the observation's times synchronization or their quantity, a significant advantage when observations are taken at different times and samples.

[GGJ09] proposes using the Bhattacharyya distance under the assumption of Gaussian distributions to evaluate the similarity between two different movements measured by a set of sensors placed on a human body. Mathematically, considering μ_{θ_i} and Σ_{θ_i} as the mean vector and covariance matrix associated with the set of values θ_i, respectively, the Bhattacharyya distance D is defined as $D = 2(1 - e^{-\alpha(\theta_i,\theta_j)})$, where $\alpha(\theta_i, \theta_j) = \frac{1}{8}(\mu_{\theta_i} - \mu_{\theta_j})'(\frac{\Sigma_{\theta_i}+\Sigma_{\theta_j}}{2})^{-1}(\mu_{\theta_i} - \mu_{\theta_j}) + \frac{1}{2}\ln\left(\frac{|\Sigma_{\theta_i}|+|\Sigma_{\theta_j}|}{\sqrt{|\Sigma_{\theta_i}||\Sigma_{\theta_j}|}}\right)$. [APM09] proposes using the Jaccard index, which does not have any a priori assumptions about the distribution of values but needs values in a small range of values: the similarity metric counts the proportion of exactly equal values returned by both observation histories. The Jaccard index can be mathematically expressed as $S = \frac{|\theta_i \cap \theta_j|}{|\theta_i \cup \theta_j|}$. Techniques for rounding observation values can be employed to increase the proportion of similar values and thus improve the effectiveness of the Jaccard index as a similarity metric.

These metrics are particularly valuable as they do not rely on observations taken at the same time. However, it is essential to note that they do not consider the spatial dimension of each observation.

3.2.2 Representing the Observation History as an Ordered Vector

One can leverage similarity metrics based on observations by interpreting them as vectors of observation values for the similarity metric. The observation histories are taken at the same time instances T, so that $\theta_i = \{\theta_{i,t}, t \in T\}$ and $\theta_j = \{\theta_{j,t}, t \in T\}$. Moreover, in general, these time instances are taken at regular intervals: $T = \{t_0 + k\tau, 1 \leq k \leq n\}$.

This type of metric is widely studied for anomaly detection [AC17; Cha09], clustering [NWN15], and classification [Zhe+20]. We list here the existing metrics related to wireless sensor observations.

Magnitude and Trend-based Dissimilarity Metric for Observation Histories

The authors of [CKJ05; LWP07; Liu+13b; TM06] define that two sensors are m-dissimilar if there is at least one time when the values of the two sensors diverge by at least m: $\exists t, |\theta_{i,t} - \theta_{j,t}| > m$. Moreover, it is suggested to combine the magnitude based metric with a trend distance, which allows for creating sensor pairs judged as similar with high

confidence. This trend metric counts the number of times that the two considered sensors have the same trends (growth or decline). Mathematically, let $\delta\theta_{i,t}$ be the rate of value increase made at time t by θ_i (respectively $\delta\theta_{j,t}$ for θ_j), such that $S = \sum_t \mathbb{1}(\delta\theta_{i,t} \times \delta\theta_{j,t} > 0)$, with $\mathbb{1}(.)$ the indicator function, that is equal to 1 if the condition is realized, 0 otherwise.

[Alm+16] suggests defining this dissimilarity as an average, so as not to misjudge sensors based on a single measurement: $S = \frac{\sum_t |\theta_{i,t} - \theta_{j,t}|}{n}$ where n is the total number of observations.

Correlation and Regression between Observation Histories

Let $\vec{\theta_i}$ denote the associated vector (with values sorted by increasing t) of θ_i.

[APM09] proposes to evaluate similarity by searching for correlation. the cosine similarity is used, which evaluates the collinearity between vectors $\vec{\theta_i}$ and $\vec{\theta_j}$: $S = \frac{\vec{\theta_i} \cdot \vec{\theta_j}}{|\vec{\theta_i}||\vec{\theta_j}|}$. [APM09; Liu+13b; Yoo20] propose using the Pearson correlation coefficient, which is based on covariance: $S = \left| \frac{\text{cov}(\vec{\theta_i}, \vec{\theta_j})}{\sigma_{\theta_i} \sigma_{\theta_j}} \right|$ where σ_{θ_i} is the standard deviation of θ_i. [KTP06] proposes to build a piecewise constant monotonic function known as an isotonic regression function to estimate the value of one sensor from the value of another.

These methods allow for highlighting hidden correlations between the sensors.

3.3 Discussion on the Choice of the Similarity Metric

Criteria for Comparing Similarity Metrics

For the analysis of the similarity measurement component, we have chosen specific points of interest for MIoT, as presented in Chapter 2 Section 5, which we specify within the context of this study on similarity metrics.

We examine whether the metric can identify particular similarities within a **complex environment**. Furthermore, in consideration of clock desynchronization issues, and the fact that the sensors are likely to have different activation periods, we take into account in the similarity computation that **assume that observations are not temporally synchronized**. Moreover, we evaluate how well the metrics handle the fact that sensors may have poor sensing hardware, which lead to **noisy observations** and the possibility of **sensor corruption**: transmission of aberrant observations.

In Table 3.2, we present a comparison of existing similarity metrics based on these criteria.

Table 3.2: Comparison of the similarity metric under MIoT criteria

Family/Proposal	References	Handling complex environment	Assuming not synchronized observations	Handling noisy observations	Identifying corrupted sensors
Geography based					
Inter-sensor distance	[Bah+14]	○	●	○	○
Based on sensing range	[SSV12; Mos+17; Mam14; KS17; CW06; TG03; AC16]	○	●	○	○
Based on statistical modeling	[Kra+06; VAA04; DB12; Raj11]	○	●	○	○
History data driven					
Bhattacharyya distance	[GGJ09]	○	●	●	●
Jaccard index	[APM09]	○	●	◖	●
Max difference and trend	[Alm+16; CKJ05; LWP07; Liu+13b; TM06]	●	○	◖	●
Mean difference, correlation, regression	[Alm+16; APM09; Liu+13b; KTP06; Yoo20]	●	○	●	●

○ Not covered: the metric does not satisfy the criterion.
◖ Partially covered: The metric addresses the criterion only partially.
● Covered: The metric fully addresses the criterion.

We have presented two main families of similarity metrics: measures based on the geographical distance between sensors and those relying on the returned observations. Now, we will discuss the advantages and drawbacks of each proposal.

Pros and Cons of Geography Based Similarity

The geography based similarities rely on distance evaluation methods, which can be highly accurate when sufficient resources are available (such as GPS), but are less accurate when relying on network information. Network metrics are known to be particularly sensitive to obstacles and are less effective in urban or vegetated environments [MF09]. Moreover, at least three gateways are required to perform position estimation, resulting in additional hardware costs.

In all models (inter-sensor distance, sensing range, statistical), it is assumed that the coverage range of a sensor is known and represented as a circular shape around the sensor. However, this simplified representation is suitable for studying simple and well-understood environments. In reality, the coverage capacity of a sensor is complex to evaluate and is not typically circular in shape. Therefore, such a measurement approach does not fully capture the complexity of real-world environments.

Another significant issue with this metric is that, since it does not rely on the observations of the sensors, it cannot consider measurement errors in the similarity computation. Moreover, it is unable to identify and differentiate a corrupted sensor that might send aberrant observations from the rest of the sensors.

One considerable advantage of this family of methods is that if the sensors do not move, the metric does not change over time, allowing for great reliability in developing an observation collection scheme.

Discussion on the Use of History-Driven Similarity

The correlation between sensor observations and geographic distance has been challenged in studies such as [KTP06; APM09], leading to a preference for data-driven similarity metrics for better reflecting the similarities between sensors. This type of metric allows for the identification of similarities in more complex environments where the similarity is not solely determined by geographic distance. For instance, grouping sensors based on the rooms they belong to can be achieved using data-driven similarity metrics, which is not possible with geography-based metrics, as developed in [LWP07].

We presented different metrics based on specific representations of an observation history. According to Table 3.2, no existing similarity metric can answer to the current challenges.

At first glance, methods based on value sets (Bhattacharyya and Jaccard) appear to fulfill the majority of the criteria we have defined. However, these metrics do not take into account the temporal aspect of observations, which limits their ability to effectively measure similarity in a real complex environment. For the Jaccard index, it is a relevant metric when the set of possible observation's values is limited, such that similar sensor transmit similar observation values. For the Bhattacharyya distance, it is assumed a Gaussian distribution of values across the sensor data, which is a relatively strong assumption.

Ordered vector-based methods share an equivalent validation of criteria with metrics based on value sets; these approaches include both value and time dimensions in the computation of the metric. However, these methods do not account for the heterogeneity of time stamps. Proposed solutions in the literature either assume an initialization phase where all sensors transmit observations simultaneously at a high rate, or rely on the unrealistic assumption that sensors transmit on the same time slots.

Open Issues

Relying on similarity measures based on observation histories is essential to explore finer similarities, not solely dependent on geography. In order to utilize a continuous similarity metric based on a sliding window, it is crucial to develop a similarity metric that can be applied to observation histories with heterogeneous sampling, taking into account both the time and value dimensions. However, we have not come across any such similarity metrics in the existing literature.

In Chapter 6, we introduce an approach that leverages interpolation techniques to handle irregular observations, subsequently defining a metric based on the average difference between interpolations. Through simulation performance evaluation, we show that metrics based on maximal differences are sensitive to noisy observations, advocating the construction of metrics founded on average differences (or correlation measures) as a preferable alternative.

Moving forward with the survey, we keep in mind that a metric based on observations must be considered. Such a metric possesses certain peculiarities: • the similarity measurement varies over time, • the similarity between two sensors may not be reliable if the observations are taken at different time instances and if the observations are noisy. These

aspects will guide our exploration of further solutions in the subsequent sections.

4 Covering Subset Algorithm

In this section, we explore methods that leverage the similarity metric to identify subsets of sensors that provide coverage.

Different Definitions of the Monitoring Objective

Across the surveyed papers, the definition of a **covering subset** varies based on the specific monitoring objectives.

In the context of the *coverage problem*, the primary aim is to guarantee adequate geographical coverage. For instance, in sensing range modeling scenarios (the predominant modeling method), the intention is to cover a maximal portion of the environment through the sensors in the covering set and their corresponding sensing areas. There is also the objective, particularly in statistical sensing models, to select a defined number of sensors in a manner that maximizes their diversity (with the term 'diversity' carrying distinct interpretations based on the employed metric).

On the other hand, for *data-driven sensor observations scheduling*, a covering subset must enable the recovery of the complete set of sensor observations with minimal measurement error. Likewise, in the context of one *sensor placement* paper under study, the objective is to select a subset from a large deployment of sensors that can evaluate the measurements of all sensors with minimal approximation error.

In less related domains, the primary objective is to create groups of sensors with strong similarity. In the *compression* field, for instance, clusters of similar sensors are formed, where only a representative per cluster transmits the compressed information to the terminal. Similarly, in the *fault detection* field, the goal is to generate clusters of similar sensors, so that if a sensor deviates from its designated cluster, an alert is raised. Here, the concept of covering subsets slightly shifts, but one can consider that having a representative sensor from each cluster contributes to forming a covering subset.

Thresholded Similarity Model

For methods focusing on *data-driven sensor observations scheduling* for instance, a threshold is applied to the similarity metric. This means that when the similarity between two

sensors surpasses a certain threshold, a link is established between them, and one sensor effectively 'covers' the other. Consequently, if a sensor is selected to be part of the covering subset, its corresponding similar sensor may not need to be actively involved, as it is already accounted for in terms of coverage.

To represent this, a graph structure can be employed, where each sensor corresponds to a vertex in the graph. Links between vertices are established if the similarity between the corresponding sensors surpasses a given threshold. Fig. 3.4 depicts a sensor similarity graph based on the similarity illustration shown in Fig. 3.1.

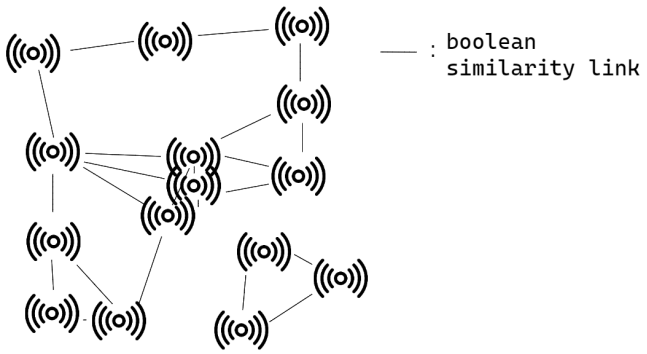


Figure 3.4: Sensors similarity graph based on the illustration shown in Fig. 3.1. A link is established between two sensors if the similarity metric exceeds the high similarity threshold (illustrated in green in the previous figure).

Covering Subset Algorithm Families

In this section, we delve into the methods proposed to address the covering subset problem. Specifically, we explore the various ways this problem is approached and categorize them into three families:

The literature offers a range of solutions that can be grouped into three main families:

- Section 4.1 - **Search for one covering sensor subset**: this approach involves identifying one covering subset of sensors.
- Section 4.2 - **Partitioning into Covering Subsets**: in this approach, the objective is to partition the sensor network into disjoint subsets, each of which is covering.
- Section 4.3 - **Clustering Similar Sensors**: this approach focuses on forming clusters of similar sensors, where one sensor is covering for the entire group of sensors within its cluster. By selecting one sensor per cluster, a covering set is established.

We provide a comprehensive presentation of these methods, including their application in the literature, as well as a discussion of any remaining open issues.

4.1 Search for a Covering Sensor Subset for Coverage Problem

One straightforward way to address the problem of reducing the number of messages is to identify one covering subset of sensors. The goal is to select the minimum number of sensors such that the set is covering.

This type of solution is proposed in the *coverage problem*. In the papers examined, sensors are selected based on their dissimilarity with other sensors. As an example, in [AC16], the objective for achieving a covering subset is to encompass 95% of the environment using sensors and their respective sensing areas.

In some papers such as [CW06; TG03; AC16; Bah+14; Mos+17], dissimilarity is based on the sensing range, and sensors are selected until the minimum required coverage is achieved. For instance, [CW06; AC16; Bah+14; Mam14] propose centralized heuristics to determine the covering subset. In [TG03], each node determines its membership in the covering subset with active neighbors; this decision is made at regular time intervals to avoid blind spots. In [Mos+17], a learning-based method is proposed where nodes are initially activated randomly, and sensors are added or removed based on evaluating the quality of the selected sensor set in terms of coverage proportion and the number of nodes selected.

On the other hand, in papers such as [DB12; Raj11], dissimilarity is based on statistical measures related to geography. In these instances, the proposed centralized methods are designed to maximize the diversity of the information transmitted to the terminal.

4.2 Partitioning into Covering Subsets

In this part, we discuss the methods that find the maximum number of disjoint subsets of sensors that are covering.

4.2.1 Partitioning the Sensor Field for the Coverage Problem

From the literature, [SP01] proposes a partitioning approach such that each subset of sensors with their respective sensing area covers the environment. A first subset is completed by adding sensors one by one until it becomes covering (from a spatial point of

view). When a subset is completed, a new one is created (if possible) using the remaining sensors. The goal is to maximize the number of subsets that each is covering.

In another approach proposed in [Kra+11], the objective is to partition sensors into subsets of a fixed maximum size. For each subset, the goal is to maximize the variance between sensors.

4.2.2 Graph-based Partitioning for Sensor Observations Scheduling

If we rely on a threshold-based measure to define pairs of similar sensors, we can establish a graph-based structure. Known graph-based methods can be applied to address the partitionning problem.

In graph theory, a subset of sensors is called dominating if, for all existing sensors, it is either a member of the set or has a common edge with it. This definition aligns with our concept of a covering subset in the context of graphs. A set of subsets are denoted as disjoint if each node belongs to at most one subset. The problem of finding disjoint dominating sets (DDS) in a graph is then to partition the sensor set into a maximum number of disjoint covering subsets.

As an illustration, in Fig. 3.5 are displayed 3 subsets that have been generated based on the graph model shown in Fig. 3.4.

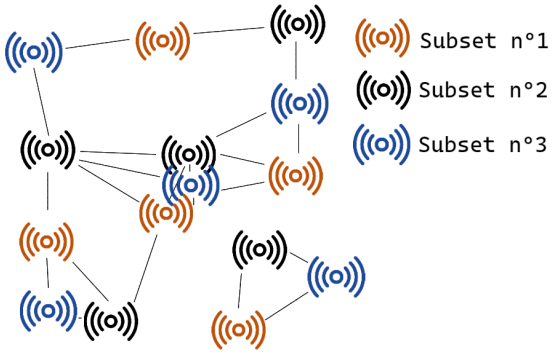


Figure 3.5: Disjoint Dominating Set generated from the sensor similarity graph in Fig. 3.4. Three disjoint subsets of sensors are displayed, and each subset is dominant.

The DDS Problem is extensively discussed in [HLR09]. This problem has been proven to be NP-complete, which implies that the resolution time increases exponentially with the size of the problem, i.e., the number of sensors. Exact resolution is achieved through constraint programming, which is defined by a set of constraints and the maximization

function, which is in this case the number of subsets. In practice, the methods proposed in the literature provide heuristic methods that find non-optimal solutions in polynomial computation time [Car+02; NH07]. In a solution proposed in [KTP06], the authors apply a DDS resolution to a directed graph, utilizing integer linear programming with the CPLEX package to determine the covering subsets.

4.3 Clustering Similar Sensors

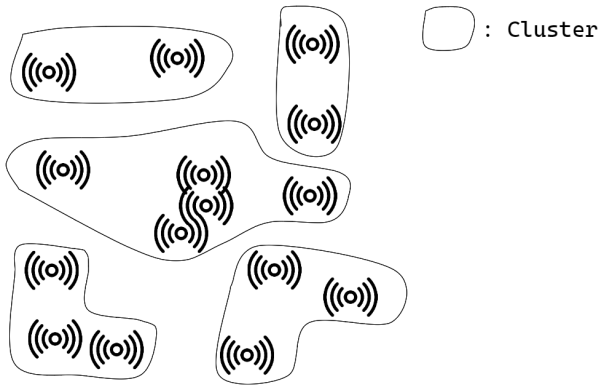


Figure 3.6: Similar sensor Clusters based on the illustration shown in Fig. 3.1. The method depicted in the figure groups sensors together based on a high level of similarity, also relaxing by allowing the grouping of sensors with medium similarity.

Clustering is a data mining technique which groups unlabeled data based on their similarities or differences[1]. Various clustering techniques are available, including hierarchical methods, partition-based methods, density-based methods, and grid-based methods [Bou07; MMR05].

In the studied case, a clustering algorithm construct groups of sensors that are considered similar. Thus, each sensor becomes covering with respect to its cluster. Therefore, selecting one sensor per cluster results in a covering set. An example of clustering result is shown in Fig. 3.6, where the clusters are created based on the similarities partially illustrated in Fig. 3.1 and the links represented in Fig. 3.4. In this example, the clustering is looking for globally similar groups of sensors and does not require full links within the clusters.

[1] https://www.ibm.com/topics/unsupervised-learning

4.3.1 Graph-based Clustering

In the literature, proposals have been made to address this problem by utilizing the concept of similar sensors based on a threshold similarity measure, transforming the problem into a graph-based formulation.

For instance, in papers related to *data-driven sensor observations scheduling* [CKJ05; LWP07; Liu+13a] or the *compression* field [TM06; Alm+16], a graph is built based on a data-driven similarity measure using the maximum magnitude and trend. The objective is to perform clique partitioning: dividing the set of sensors into groups in such a way that within each group, every pair of sensors are connected; this is a known NP-complete problem. To solve this, heuristic approaches have been developed, like a greedy approach in [LWP07] that builds a cluster by iteratively aggregating the available fully linked sensors, starting by the largest degree nodes.

4.3.2 Clustering for Fault Detection

Clustering methods for fault detection in sensor networks are also found in the literature. The goal of these methods is to group sensors that make similar observations to detect when a sensor reports observations that deviate from the norm of the cluster, as presented in the survey [PSP22]. For instance, in [Yoo20], a Gaussian mixture model is used to cluster sensors based on observation driven similarity between pairs of sensors. Each cluster defines a normal behavior framework for all the sensors it contains: a fault is detected if a sensor's observations fall outside the bounds of its cluster. In this method, groups of similar sensors are created, but this is not interpreted in terms of covering sets.

4.3.3 Clustering for Sensor Placement

By deploying a large number of sensors for a testing phase, it becomes possible to apply clustering methods to identify key locations for the placement of a reduced number of sensors. Such an approach is proposed in [Yog+18], where a clustering method is used to group sensors based on their data-driven characteristics. The center of each cluster is then selected as a reference point, allowing the replacement of the entire cluster with a single sensor positioned near its center. The clustering process is conducted in two steps. Initially, the k-means clustering algorithm is employed with a data-driven approach. Subsequently, the density-based clustering algorithm is utilized, incorporating a geography-based metric to distinguish clusters that exhibit similarity in data but are

physically distant from each other. This methodology aids in optimizing the sensor deployment and reducing the number of sensors required while preserving critical monitoring locations. Here, the selected sensors can be seen as a chosen covering subset.

4.4 Discussion on the Choice of the Covering Subset Algorithm

Chosen Criteria for Covering Subset Algorithms Comparison

We now proceed to compare the various solutions aimed at looking for covering subset(s). In Table 3.3, we evaluate these solutions based on some points of interest outlined in Chapter 2 Section 5.

We evaluate whether the solution is capable of **utilizing the entirety of the sensors**. As we will delve into in the subsequent section (Section 5.3), relying on only a subset of sensors to fulfill the monitoring requirements is not ideal. Specifically, receiving observations solely from dissimilar sensors makes the detection of corrupted sensors impossible. Additionally, we will see that in practical implementation, utilizing only a portion of sensors introduces additional costs at the sensor level.

We assess their adaptability to different scenarios, taking into account variations in the sensor field, including the **management of adding or removing sensors**. In terms of handling changes in the environment and the variability of measured similarity links, we focus on scenarios involving the **adaptation to changes in similarity links**: cutting a similarity link, adding a link between sensors or groups of sensors. These two criteria are evaluated based on the extent of modification required in the solution (in terms of modification of the covering subsets) in response to these perturbations. It's important to note that substantial alterations to the covering subsets inevitably entail significant management costs, i.e., need for instructions from the terminal to the sensors.

Lastly, we will investigate whether the methods **address the issue of unreliability in similarity measurements between sensors**: to what extent the measurement errors of sensors negatively affect the effectiveness of covering subset search.

Here, we discuss proposed methods as resolution models. We do not account for the constraints outlined in the studied papers. As an example, for methods relying on geography-based similarity, we investigate the robustness of the covering algorithm to variations in the similarity metric over time. This approach is realistic since sensors could be in motion, even though such scenarios are not addressed in the analyzed papers.

Table 3.3: Comparison of the covering subset algorithms under MIoT criteria

Family/Proposal	References	Utilizing the entirety of the sensors	Managing input and output of sensors	Adapting to changes in similarities	Addressing unreliable similarities
Search for one covering sensor subset					
Heuristic	[CW06; AC16; DB12; Raj11; Bah+14; Mam14]	○	●	◖	⌀
Set-by-step	[TG03; Mos+17]	○	●	◖	⌀
Partitioning into covering subsets					
Heuristic	[SP01; Kra+11]	●	◖	◖	⌀
DDS	[KTP06]	●	◖	◖	◖
Clustering similar sensors					
Clique partitioning	[CKJ05; LWP07; Liu+13a; TM06; Alm+16]	●	●	●	◖
Traditional clustering	[Yoo20; Yog+18]	●	●	●	●

⌀ Not applicable: it is not possible to evaluate the criteria for the proposal.
○ Not covered: the proposal does not satisfy the criteria.
◖ Partially covered: The proposal addresses the criterion only partially.
● Covered: the proposal meets the criteria.

Critique of Solutions Based on One Covering Subset

Solutions based on a single covering subset generally exhibit good adaptability to the uncertainties of MIoT. In most cases, minimal modifications to the covering subset are required when a link is added or removed, or when a sensor is added or withdrawn. However, a major drawback of this approach is that it relies on only a portion of the available sensor pool. As elucidated in the presentation of the criteria, this choice has limitations, as it results in additional sensor consumption and prevents the identification of a corrupted sensor once it is in use.

Critique of Solutions Based on Partitionning Covering Subsets

Solutions relying on multiple covering subsets encounter issues related to their inflexibility. When changes occur in similarity links, such as the addition or removal of a link, it typically necessitates the search for entirely new covering subsets. Similarly, when a sensor is added or removed, a completely new configuration of the covering subsets needs to be devised, and modifying the existing solution to accommodate these changes is not straightforward.

In cases of perturbations within the MIoT, these solutions must be recreated from scratch. Generally, the new covering subsets are quite different, which implies significant modifications to the subsets.

Discussion on the Development of the Clustering Method

In contrast to methods based on finding one or multiple covering subsets, the methods based on the search for similar sensor groups are inherently more flexible.

By assuming an existing structure of similar clusters, when a new sensor is added, the clustering method will choose to include the sensor in one of the existing groups. Similarly, if a sensor becomes inactive, it is simply removed from its cluster.

When there are variations in similarities, clustering methods are likely to require few modifications to their structures. For instance, in the case of clique partitioning, the heuristic initially groups together sensors with the highest similarity rates. Adding or removing a link subsequently results in only slight alterations to the order of sensor joining. More generally, it is conceivable that the emergence of new similarities lead to the merging of two clusters; if similarities disappear, the connected cluster may separate into multiple clusters.

An additional significant advantage, not accounted for in the criteria-based comparison in Table 3.3, is the ability to adapt to the variability in the number of similarities. Clustering methods can effectively isolate poorly placed sensors in individual clusters, create smaller groups for localized phenomena, and form larger clusters for more prevalent phenomena. This level of flexibility is not possible in partitioning-based methods.

Open Issues

Based on this review, we can conclude clustering methods appear to be a preferable choice, as they demonstrate superior flexibility, adaptability to variations in similarities, and the capability to handle the variability in the number of similar sensors.

In Chapter 6, we introduce a hierarchical clustering approach designed to group sensors based on their overall similarities. We compare our conventional clustering method with a threshold-based clustering approach, which, as demonstrated in this section, serves as the primary reference. Through our simulations, we demonstrate that methods that rely on setting a threshold on metrics computed between pairs of sensors, especially when dealing with noisy observations and varying time instances, tend to underperform due to their inflexibility. Our findings suggest that embracing conventional clustering methods is more effective in managing such uncertainties.

5 Activation Allocation Method

In this section, we study what proposals are made in the literature to define explicitly the output of the observation collection scheme. An activation allocation method determines the modes of the sensors over time: sleep mode, where the sensor is inactive in monitoring, and active mode, where the sensor observes the environment and sends observations to the terminal.

Based on our analysis of the literature, we have identified two main families of activation allocation methods, which we will elaborate on in the following subsections:

- Section 5.1 - **One covering set in active mode**: Only one covering set is utilized at a time, while the other sensors remain in deep sleep mode.

- Section 5.2 - **Round-robin**: Sensors are activated in a manner that forms a round-robin sequence among the different covering subsets.

The choice of the activation allocation method depends on the output of the covering subset algorithm, which can have three different representations. If a covering subset is selected, then it is not feasible to use methods based on round-robin, and it is proposed in the literature to use this covering set as the active one, the others being in deep sleep. On the other hand, when the decision is to use a partitioning into covering subsets or clustering of similar sensors, the literature consistently opts for an activation allocation method based on round-robin, although it is technically feasible to select one covering subset as the active one.

We will examine the solutions proposed in the literature, and we will discuss the most relevant methods for our MIoT scenario.

5.1 One Covering Set in Active Mode

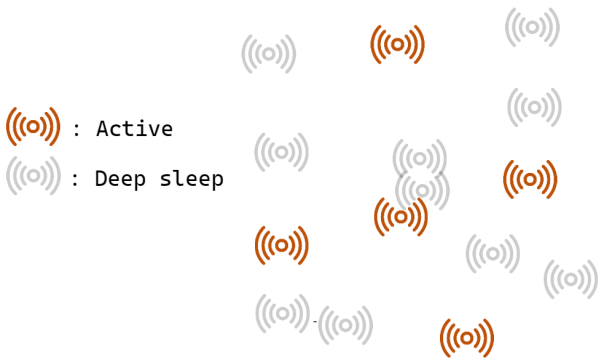


Figure 3.7: Activation allocation method where always the same active sensor subset is chosen. The figure is based on the graph-based DDS resolution shown in Fig. 3.5, with the active subset corresponding to one of the covering sets. The sensors in the chosen covering subset activate at regular intervals, while all other subsets of sensors remain in deep sleep mode.

One approach to defining the activation allocation method is to select a covering subset of sensors that will be activated while the remaining sensors stay in deep sleep mode. The covering subset can be modified over time. For example, as illustrated in Fig. 3.7, the Disjoint Dominating Set (DDS) resolution from Fig. 3.5 can be used for this purpose.

5.1.1 Same Active Subset Until a Sensor Fails

In the conventional approach, once a covering subset is selected, [Raj11; AC16; Bah+14; DB12; Mos+17] propose to consistently use this subset of sensors in active mode until a

sensor failure occurs.

Role of Active Sensors

In [Raj11; Bah+14; DB12], the activation mechanism of the sensors is not explicitly specified, implying that the activated sensors remain constant over time. In [AC16], two types of sensors are considered: trigger-based sensors that transmit messages upon detecting changes, and periodic sensors that activate at regular intervals (without specifying the period).

In [Mos+17], a two-phase process is introduced. During the initial phase, the active subset of sensors is updated at each time step. Upon completing this phase, the final solution is derived, and the set of active sensors remains unchanged throughout the operation.

Take Over when a Sensor Fails

In the event of a sensor failure, other sensors need to be in active mode to assume the responsibilities. Two approaches are proposed for transferring responsibility when a sensor fails. The first method involves maintaining the inactive sensor in a listening mode, awaiting a handover directive [AC16; Raj11]. While ensuring swift handover, this solution demands substantial energy due to continuous radio activation.

The second method periodically switches sleeping sensors to listening mode, gradually shortening the listening period using a Weibull distribution, as outlined in [Bah+14]. As time advances, the likelihood of a sleeping sensor taking over from a defunct one increases, leading to a reduction in its listening mode duration. This second approach is more realistic in terms of energy consumption, yet it might introduce latency in the handover process as the sensor needs to transition to the listening mode before becoming active.

5.1.2 Regular Updates of the Active Sensor Set

Another approach is to regularly change the active subset of sensors. In [TG03], at each round, a sensor decides whether to activate based on its similarities with its neighbors. The activated sensors stay active and transmit every 0.5 seconds during the entire round duration of 10 seconds.

In [Mam14], at each round, the active subset is chosen based on various criteria such as the number of neighbors, shared sensing region, residual energy, and the number of consecutive times the node has been in active mode. The selected subset remains active for the entire duration of a round. The objective here is to avoid excessive use of a particular sensor and to balance the energy consumption among all sensors.

5.2 Round-Robin

Another activation allocation method is based on the slotted timeline and round-robin methods applied to subsets of sensors. In this model, sensors are activated in a way that, in turns, covering subsets becomes active in a round-robin fashion.

5.2.1 Round-robin between Disjoint Covering Subsets

After using a method based on partitioning into disjoint covering subsets, several studies, such as [KTP06; Kra+11; SP01], propose applying round-robin between each covering subset. In this approach, during each round, one covering subset becomes active while the rest of the sensors remain in deep sleep mode. In [KTP06] and [Kra+11], the covering subsets are activated in a round-robin fashion every 30 seconds, while [SP01] does not specify the duration of a round.

5.2.2 Round-robin within Each Cluster of Similar Sensors

Starting with clusters of similar sensors, [CKJ05] proposes load balancing within each cluster. The sensors belonging to a cluster are activated in a round-robin fashion to receive messages at regular intervals, for example, every 5 minutes. A revised version of [CKJ05] is proposed in [LWP07]. In this new approach, sensors no longer have fixed transmission periods; instead, they are assigned activation probabilities for each time slot. The solution introduces a factor to increase the likelihood of multiple sensors in the same cluster transmitting simultaneously. This modification aims to improve packet loss management and enables the evaluation of the ongoing relevance of similarities.

Another probabilistic activation approach is proposed in [Liu+13a], where a transmission probability is assigned to each node in a cluster, with all sensors transmitting at similar time stamps. The activation probability is inversely proportional to the cluster size and linearly dependent on the distance from the terminal. As a result, this method achieves a balanced distribution of transmissions among the sensors within a cluster.

Moreover, sensors located closer to the terminal are activated more frequently, as they play a crucial role in relaying observations within the mesh network.

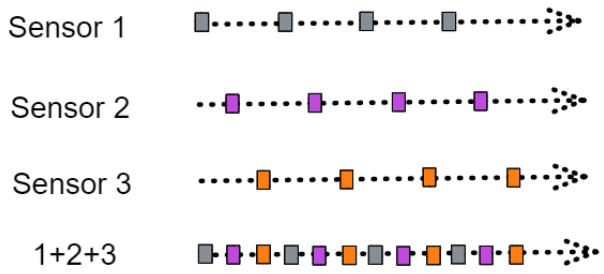


Figure 3.8: Three sensors transmitting in a round-robin manner: each sensor's observation is represented by colored squares. The result shows the set of observations made by the three sensors: receptions of observations at regular intervals, evenly distributed among the three sensors.

Fig. 3.8 illustrates the round-robin method applied to a set of 3 sensors belonging to the same cluster, ensuring the reception of messages from the cluster at regular time intervals.

5.2.3 Translation of a Round-Robin Method into Activation Period Updates

In the papers we have presented, only theoretical round-robin principles have been exposed. We propose a translation of a round-robin method into the form of activation period updates.

The sensors operating in a round-robin fashion share a common activation period, which is a multiple of the number of rounds within the round-robin cycle. This period is fixed, and in case of adding or removing rounds, activation periods need to be redefined, during the active mode of the concerned sensors.

Moreover, when a sensor enters the environment, it needs to be included in the scheduling time steps. According to our communication hypothesis, the new sensor must receive two consecutive orders to adjust its activation period and be included in the scheduling time steps.

5.3 Discussion on the Choice of the Activation Allocation Method

Chosen Criteria for Activation Allocation Method Comparison

We have presented activation allocation families that we propose to compare in Table 3.4. We rely on the points of interest outlined in Chapter 2 Section 5 to highlight the different aspects that an activation allocation method should fulfill. We assess whether the solution **minimizes the control of sensors**, which pertains to the reduction of listening times for inactive sensors, directly impacting their battery longevity. Then, we evaluate whether the distribution of activations **allows for the identification of corrupt sensors**. Finally, we **assess the flexibility of the solution concerning clock drift**.

Table 3.4: Comparison of the activation allocation methods under MIoT criteria

Method	References	Minimizing the control of sensors	Identifying corrupt sensors	Adapting to clock drift
One covering set in active mode				
Same active set	[Raj11; AC16; Bah+14; DB12]	○	○	●
Updates of the active set	[TG03; Mam14]	○	◖	○
Round-robin				
Between disjoint covering subsets	[KTP06; Kra+11; SP01]	●	●	○
Within each group of similar sensors	[CKJ05; Liu+13a; LWP07]	●	●	○

○ Not covered: the proposal does not satisfy the criteria.
◖ Partially covered: The proposal addresses the criterion only partially.
● Covered: the proposal meets the criteria.

Critique of the Activation of One Covering Subset

We have identified significant limitations to the activation of a covering subset of sensors. The first lies in the power consumption of the sensors that are not actively used in monitoring. In order to take over, these sensors must periodically enter a listening mode. This consumption can become very significant when a large majority of deployed sensors are not used in the active covering set, but still need to consume energy during the listening mode.

The other limitation is related to the lack of information diversity. Solutions based on one covering subset heavily rely on the active sensors. Consequently, in the event of an issue with an active sensor (e.g., a corrupted sensor), it is challenging to identify the problematic sensor since there are no other active similar sensors.

Furthermore, in solutions that involve periodic handover (when the active subset is updated regularly), all sensors must be synchronized to enter listening mode at the same time during the change of the active sensor subset. This kind of choice is complex to implement, especially for low-cost sensors considered, sensitive to clock drift.

Discussion on the Round-Robin Model

The round-robin method addresses most of the limitations of solutions based on using only one active covering subset. In this approach, all sensors are utilized in the monitoring process. Since similar sensors take turns in activating, it becomes possible to identify if one of them deviates from normal behavior. Additionally, this type of solution does not lead to overconsumption of sensors as long as there are no variations in similarities or in the number of sensors.

Nonetheless, these methods rely on activation synchronization, a challenge discussed in the preceding section, which is difficult to achieve in practical scenarios. Sensors are susceptible to clock drift, leading to slight variations in their actual activation periods. Consequently, frequent readjustments are necessary to maintain synchronization.

Another aspect worth mentioning, although not indicated in the criteria of Table 3.4, is that to include a new sensor, it needs to synchronize with the existing time steps. For this to happen, the new sensor must receive two consecutive orders. This can be particularly problematic if transmissions are not guaranteed, and receiving two consecutive orders from the terminal might be hard to do.

Open Issues

We have presented two main families of activation allocation methods. However, none of the proposed solutions in the literature fully satisfies the criteria we have defined for an MIoT deployment. Among the available options, the most feasible approach seems to be the round-robin-based methods, which involve activating all sensors in the network. Nevertheless, the existing solutions require strict synchronization among sensors to ensure that estimation updates occur at strictly regular intervals. This constraint poses a significant limitation to the effectiveness of the solution.

In Chapter 4, we provide a formal definition of an activation allocation method as a function that redefines the activation period of a newly activated sensor, and we propose a technique that activates a predefined maximum number of sensors using a round-robin approach. The approach outlined in Chapter 5 relaxes the strict message reception requirement and enables a more equitable distribution of the workload among all sensors within similar clusters, achieving a significantly reduced number of period changes.

6 Conclusion

In this survey, we have explored methods for reducing sensor transmissions in a dense sensor deployment while considering the constraints of a MIoT deployment. To achieve a reduction in observation transmissions, an appropriate approach is to leverage the principle of similarity. This involves three core components: the **similarity metric**, the **covering subset algorithm**, and the **activation allocation method**. These components have been thoroughly studied and characterized, taking into account factors such as a heterogeneous environment, low-cost sensors, and network constraints.

Regarding the first component, discussed in Section 3, we have concluded that establishing similarity based on sensor observations is highly beneficial for effectively managing complex environments. However, current solutions lack the ability to incorporate similarity based on observations taken at different time instances.

Moving on to the second component, addressed in Section 4, we have highlighted the advantages of using clustering approaches for similar sensors, as they enable better adaptation to system dynamics.

Finally, in the third component, presented in Section 5, we have found that round-robin methods are the most viable solution for reducing sensor power consumption and identifying corrupted sensors. Presently, round-robin methods rely on a strict scheduling approach, limiting its real application in practice.

For the subsequent chapters constituting the contribution of this manuscript, we make choices grounded in the conclusions drawn from this survey. We outline the overarching scheme as follows: we employ a similarity metric based on sensor observations, acknowledging its associated uncertainties and potential temporal fluctuations. With this similarity metric as a foundation, we partition the sensor set into clusters of similar sensors. Given the dynamic nature of these groups, owing to sensors entering and exiting clusters and evolving similarities, we propose a method for distributing the observation load

within each cluster.

Part II

ACTIVATION ALLOCATION METHOD

Chapter 4

Selecting a Subset of Sensors in Round-Robin

In the preceding section, we introduced the framework of our study, which we presented within the context of a Massive IoT paradigm. We emphasized the significance of developing effective observation collection methods based on similarity, a topic we then extensively explored through a literature review.

In this chapter, we formally define the period allocation function component, as updates of the activation period of a sensor that has just activated, according to the knowledge up to that time. We then propose a solution that allows having a constant number of messages returned by a whole sensor fleet. Without prior knowledge of the number of sensors entering the environment dynamically, the proposed method guarantees regular and strict interval reception of observations, with a predefined maximum number of sensors activating in round-robin. Such a solution can be seamlessly integrated to distribute the observation load within a cluster of similar sensors.

Our experimental results highlight the trade-off between tracking quality and system lifetime. Moreover, we demonstrate that when dealing with a relatively large sensor fleet (e.g., 300 sensors), it is advantageous to select only a subset of sensors to activate in a round-robin fashion. Moreover, the target activation period, a parameter of our period update function, also has a significant impact on this trade-off between tracking quality and system lifetime.

1 Problem Statement and Model

1.1 Assumptions and Notations on Sensors

Throughout this thesis, we are interested in the monitoring of an environment with IoT sensors. Sensors are dynamically integrated in the management system at the time of

their first activation, also called **initialization**, that we denote by t_i for the sensor i.

Considering the already presented sensors transmissions modeling, recall that a sensor is in activation mode periodically, transmitting a message, and during a short period of time after the transmission, it is possible to modify the activation period of the sending sensor. As the main objectives of this manuscript revolve around managing sensors activations, we develop the **period allocation** as a function f, defined in more detail hereafter, for that purpose.

A sensor is said to be **alive** at a time t if, at that time, its initialization is passed and it has enough energy to activate again. Conversely, a sensor that is not alive anymore at time t is said to be **dead**. We talk about the **end of monitoring** when there are no more alive sensors. For the whole chapter, we consider proposals where the n sensors come alive without the monitoring stopping in the meantime.

We consider known the initial energy of a sensor i, equal to e_i. In the strategy presented hereafter, we take into account a consumption model based on radio energy consumption, as it is the most significant factor compared to other sources of energy consumption [Bou+18]. We denote the sensor energy consumption for each sensor-to-terminal observation activation by c_e, and the sensor energy consumption for the terminal-to-sensor period change orders by c_r. We assume that each sensor utilizes its energy until its battery is depleted.

1.2 Formalization of the Period Allocation Function

In Chapter 3, we have discussed on the various components necessary for the development of an active mode control policy: a similarity metric, a covering subset algorithm, and a period allocation. In this section, we focus on formalizing the period allocation function in the context of constraint network and constraint sensors. Here, the function is responsible for redefining the activation period of a sensor upon receiving a sensor message. The function takes the sensors' knowledge history up until that point as an input and returns a new activation period for the sensor.

Definition 1. *Let us denote by H_t the knowledge up to and including time t.*

*A **period allocation function** is a function f:*

$$f : H_t \to \mathbb{R}^{+*}, \tag{4.1}$$

where $f(H_t)$ represents the new activation period for a sensor that has just sent a message

at time t.

The function f is used for each new received message. In particular, f defines the initial period of sensors. If the function f returns a different period from the sensor's current one, a downlink activation from the terminal (with an energy cost c_r to the sensor) takes place to modify that period, so that after sending a message at time t, a sensor's period always equals $f(H_t)$.

1.3 Ensuring Regular Observations from a Sufficient Number of Sensors

The objective of this chapter is to study the influence of the period allocation function on tracking quality and sensor consumption. By using the three-component architecture developed in the state of the art, let us consider a scenario where a cluster of similar sensors has been identified. Our aim is to limit the overall number of messages sent by that cluster by adapting to the number of sensors present.

Nevertheless, solely receiving observations from a limited subset of sensors might not provide a comprehensive and reliable view of the environment. Indeed, it remains crucial to gather observations from diverse sources to detect variations in sensor similarities, for instance identify potential sensor corruption.

Additionally, we will investigate the impact of the update period on the accuracy of the estimations. Hence, our goal is to study the trade-off between the number of sensors to transmit to update estimations with defined period, and the quality of tracking.

1.4 Definition of the Quality Metric

To quantify the quality of the messages received by the terminal, we introduce the concept of **diversity** which measures the amount of information received from various sources weighted by their relative importance.

In this context, we define the relevance of a piece of data based on its aging. The **freshness** of a message evaluated at t [Bou04; ES07; SC19] represents the relevance of the activated information as a function of its age. This is a positive decreasing function taking as argument the difference between the observation time t and the message sending time t', i.e., $\Delta_t = t - t' > 0$.

Sensors send messages to the management system, updating their activation period after an activation if told so by the terminal. We apply the notion of freshness to a sensor

by considering its most recent activation, in order to propose the following definition of diversity. Illustrations depicting the freshness evolution of two sensors over time are provided in Fig. 4.1(a).

Definition 2. *The diversity at time t is defined as the sum of the freshnesses of all sensors that are or were alive at that time.*

*The **average diversity** is the average of the diversities over the entire monitoring duration. The average diversity related to a period allocation function f is denoted by $D(f)$.*

Below are two examples of freshness functions:

- $u_T(\Delta_t) = \mathbb{1}_{\Delta_t<T}$, for some value $T > 0$, meaning that the value of some received data remains constant during T then suddenly drops to 0.

- $v_T(\Delta t) = \exp\left(-\frac{\Delta t}{T}\right)$, with a smoother depletion of the information value over time.

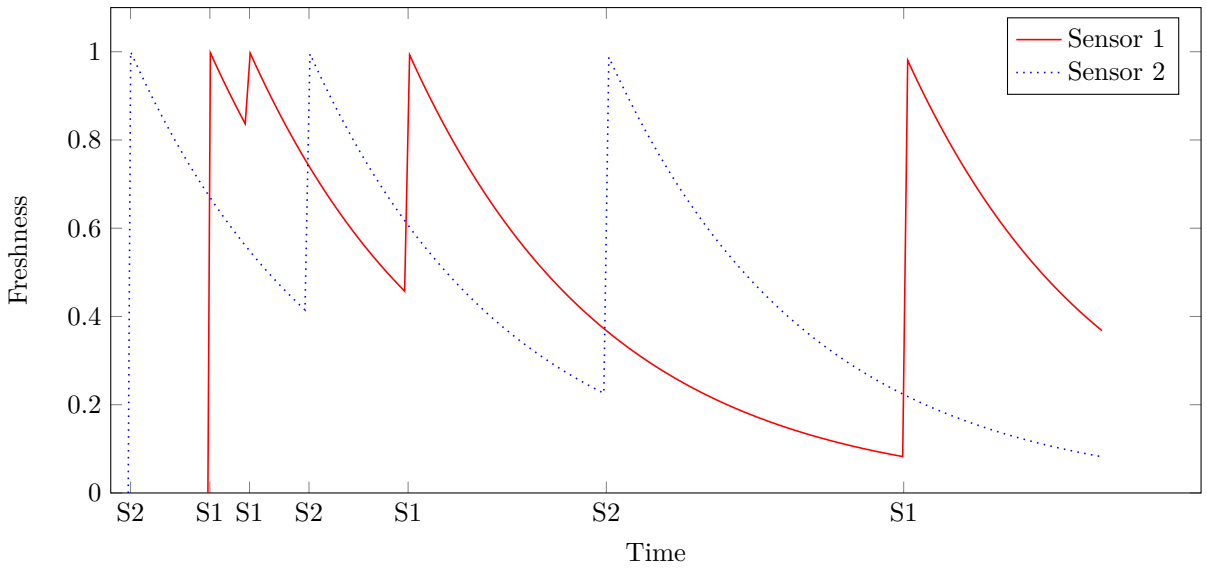
The parameter T characterizes the relevance time of data: if T is large, then we consider that "old" data remains relevant.

To illustrate the meaning of the diversity measure, consider the freshness function $u_T(.)$: if a period allocation function f induces a diversity X, then that means that over a sliding window of size T, messages are received on average by X different sensors. Similarly, the diversity measure depicted in Fig. 4.1(b) using the function $v_T(.)$ illustrates an example with two sensors.

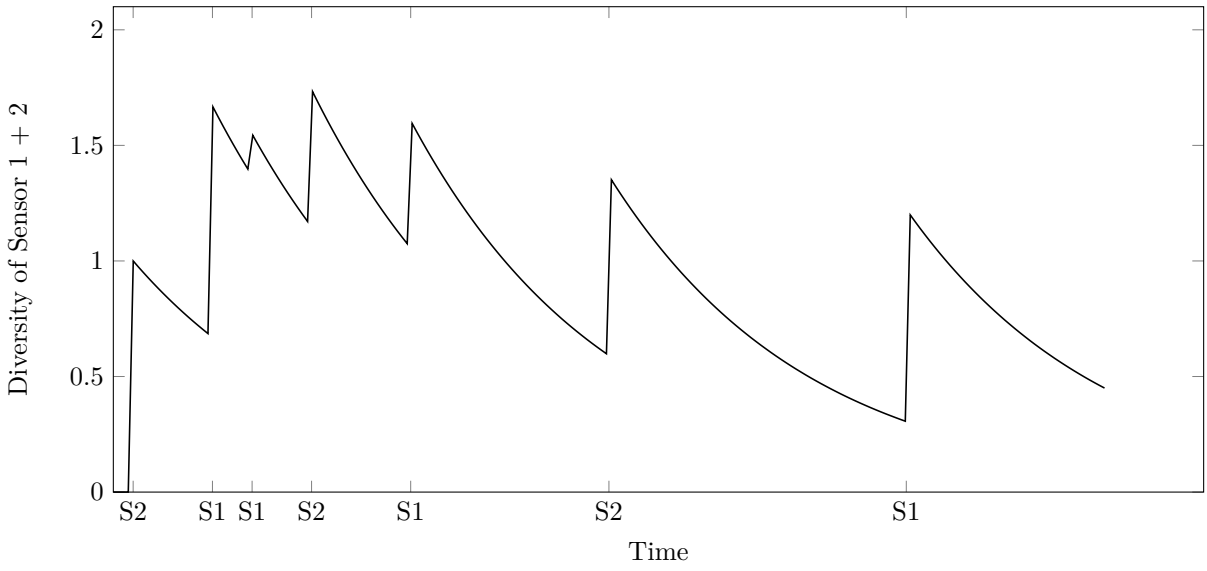
1.5 Definition of the Monitoring Duration

We utilize a comprehensive energy efficiency metric known as the **monitoring duration**. This metric represents the total duration of the monitoring period, starting from the initialization of the first sensor and extending until the end of the monitoring (death of the last).

In this way, we characterize a bi-objective problem: we can quantify and compare the qualities of period allocation functions through our two performance metrics, for energy efficiency and monitoring quality.



(a)



(b)

Figure 4.1: Evolution of freshness over time for 2 sensors in (a) and the corresponding diversity in (b). The x-axis represents time, with the observations sent by the sensors labeled as $S1$ and $S2$. The y-axis represents the freshness (or the sum of freshness) according to the observations, with the exponential freshness function.

2 Ensuring Periodic Activations From at Most M Sensors

In this section, we develop a strategy to guarantee, by defining the period allocation function, that there is one (and only one) periodic activation, with a period τ, and that at most M sensors activate in turn (M and τ are chosen by the monitoring manager).

2.1 Definition of Effectiveness

As stated in the initial problem statement (Chapter 2), our objective is to track the average value of a physical quantity over time by gathering observations at regular intervals.

Hence, starting from the instant of the first message received at time t_0, we want in this part to receive exactly one message at regular time intervals from one of the alive sensors. Referring back to the term **target activation period**, denoted here as τ, this characteristic of strictly periodic reception will be referred to as **effectiveness**. This property is formalized as follows:

Definition 3. *A period allocation function is said to be* ***effective over the instants of period*** τ *if the sensor activations verify that:*

- *Starting at* t_0*, one and exactly one activation is made at each target activation period* τ *as long as there are alive sensors.*
- *Apart from the initialization, no sensor activates between each time interval* τ*.*

Consider Π as an initial scenario: a sensor $i \in \Pi$ has an initialization time t_i and an initial energy level e_i. The size of the sensor set is denoted by $|\Pi| = n$.

For this set Π, and considering an effective period allocation function over time intervals of duration τ, we can quantify its efficiency by calculating the sample span.

Definition 4. *Given an effective period allocation function* f *and a sensor set* Π*, its* ***sample span*** $L(f, \Pi)$ *is defined as the number of consecutive activations over the instants of period* τ *until the end of the monitoring.*

The **monitoring duration** of a period allocation function effective over the instants of period τ is then simply defined as $\tau L(f, \Pi)$.

We develop below an analytical upper bound for the span (and thus, duration) of effective period allocation functions.

Proposition 1. *For an effective period allocation function f and a sensor configuration Π of size n satisfying the condition that all sensors initialize without the monitoring stopping in the middle and that no sensor comes alive exactly at an instant of the form $t_0 + k\tau$ for an integer k, the sample span is upper-bounded by the following expression:*

$$L(f, \Pi) \leq \frac{(\sum_{i \in \Pi} e_i) - nc_e - (2n-1)c_r}{c_e} \tag{4.2}$$

The proof is developed in Appendix B.1. This result is based on the observation that in order to achieve an effective allocation function, we need to change the activation period of sensors (except the first one) at least twice.

In the rest of this section, we develop a specific function, that we will denote by $f_{M,\tau}$, and that we will show is effective over the instants of period τ, while jointly using up to M sensors to provide some diversity.

2.2 Overall Principle of the Period Allocation Function

Considering the scenario of tracking an average physical quantity, we want to develop a period update function allowing to receive messages at regular intervals; the target activation period τ is the first parameter of our function.

In a context such as the one predicted for MIoT, it is possible to have faulty sensors or variations in the similarities among sensors. It may then be necessary to receive information from various sources (quantified by the average diversity). The second parameter of the function, that we will denote by M, will be the number of sensors activating in a round-robin fashion. The selection of this value for M influences the frequency of necessary period changes; a higher count of sensors in round-robin entails more frequent period adjustments.

For given parameters τ and M, we therefore want to define a period allocation function $f_{M,\tau}$ such that at most M sensors activate in turn, with periodic activations of period τ.

If any, the other sensors will be set in sleep mode, and successively take over the dead sensors.

• When the number of alive sensors is below M, all alive sensors activate in turn. In that case, each alive sensor has an activation period set to τ times the number of alive sensors. As long as the number of sensors is below M, if a new sensor comes alive or dies, all the sensors then change their activation period to maintain that property.

• As soon as the number of alive sensors is more than M alive sensors, our proposed scheme works differently: M sensors activate periodically, with a period of $M\tau$, and the period of all the other sensors is set so that they successively take over dead sensors. When one such sensor takes over the death of another one, its period is set to $M\tau$, to ensure the same role.

An illustrative example of sensor activations using the period allocation function is shown in Fig. 4.2. The activations of each sensor i, are represented as dots along the horizontal line corresponding to $y = i$. The black squares indicate period changes after each activation, as determined by the function $f_{M,\tau}$.

As per the initial objective, there is precisely one activation on each target activation period τ, except for the first activation. This can be observed on the upper horizontal line, which aggregates all the activations.

It is worth noting that when there are at least 3 sensors active simultaneously, the activations are cyclically shared among these 3 sensors, following the periodic activation pattern established by $f_{M,\tau}$.

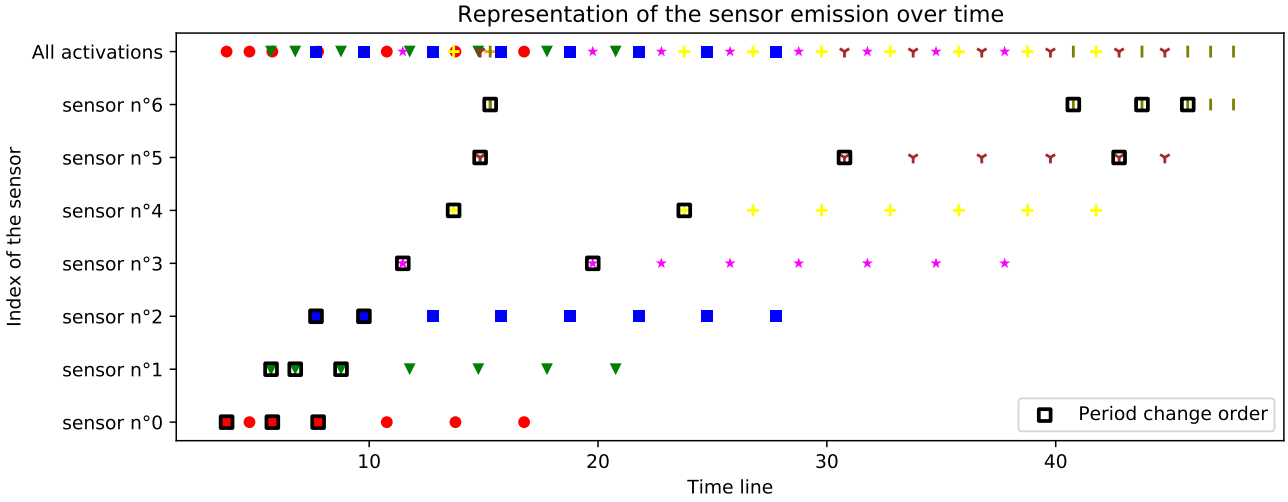


Figure 4.2: Illustration of sensor activations using the function $f_{M,\tau}$ with parameters $M = 3$ and $\tau = 1$. In this scenario, we have a set of 7 sensors, each with equal battery capacities ($e = 10$) and activation and period change consumption ($c_e = c_r = 1$). These sensors become active at random times between $t = 0$ and $t = 25$.

2.3 Formal Definition of the Period Allocation Function

The function is constructed in two parts, distinguishing between the first activation of the sensor and the case where it is already alive. We denote by $|\Pi(t)|$ the number of sensors at time t, including a new sensor initializing at t, or deleting a sensor whose last activation took place at this time.

• First, the activating sensor was already alive before. When the total number of alive sensors does not exceed M (i.e. an insufficient number of sensors has come alive or a too important number of sensors is out of battery), the function defines for each sensor a period $f(H_t) = |\Pi(t)|\tau$.

When there are enough alive sensors ($|\Pi(.)| > M$), then $f(H_t) = M\tau$. Thus, for already alive sensors, the period allocation function is defined by:

$$f_{M,\tau}(H_t) = \min(M, |\Pi(t)|)$$

• When a sensor comes alive, there are 2 different cases:

· If the number of alive sensors is less than M, then new sensors get included in the round-robin scheduling. When a sensor initializes, it needs to be scheduled over the instants of period τ, activating τ units of time after the sensor that activated just before. $(t - t_0)\%\tau$ (where "%" represents the remainder operator in the division algorithm) represents the time between the previous time interval of length τ and t. Since $\Pi(t)$ has been incremented by one upon the arrival of this new sensor, the period allocation function is defined by:

$$f_{M,\tau}(H_t) = \tau|\Pi(t)| - (t - t_0)\%\tau \tag{4.3}$$

This last formula also works in the case where multiple sensors are initializing between two instants of period τ.

This principle is illustrated in Fig. 4.2, where the example of sensor number 2 is depicted. In this case, the activation period of sensor 2 (blue square) is adjusted after its first activation so that it transmits after the first two sensors that were already active. Following the initialization of sensor 2, the activation periods of sensors 0 and 1 are adjusted to include sensor 2, with their new activation period of 3τ.

· We consider now the second case - one sensor initializing when there are already at least M alive sensors. Then, a sensor that comes alive is put to sleep until a sensor dies. The sensor takes over from a sensor whose relay is not taken, i.e., activates $M\tau$ after its last

activation:

$$f_{M,\tau}(H_t) = \text{death time of a sensor } -t + M\tau$$

We now introduce an object allowing to keep in memory the predicted deaths of each sensor. We define *death-date* the list sorted by ascending date of the dead sensors whose death is not covered by a relaying sensor. Algorithm 1 defines the **death update algorithm**, updating the list *death-date*. This algorithm is executed each time a message is received, following the use of the period allocation function. When a new sensor initializes, we use the first element of the death-date list to determine its initial activation period. Since the relay for this sensor has been assured, the corresponding element is removed from the list.

Certain specific aspects of this algorithm have been partly commented upon in the pseudo code, enclosed within curly braces. Additionally, we provide detailed explanations for lines 5 and 7. The purpose of these lines is to update the "death date" of sensors in the scenario where a sufficient number of sensors initialize, i.e., $|\Pi(t)| > M$. In this scenario, if a sensor previously had a period of $M\tau$, from its next activation onwards, it consumes the remaining energy $(e_i - c_e)$ to transmit at a period of $M\tau$. Conversely, if the condition is not met, the sensor must adjust its period during its next activation to $M\tau$, resulting in a remaining energy of $e_i - c_e - c_r$ after the next activation. From that moment (which is $t + p_i$), it will transmit all its energy at intervals of $M\tau$, equivalent to a duration of $M\tau \left\lfloor \frac{e_i - c_e}{c_e} \right\rfloor$ if $p_i = M\tau$ (or $M\tau \left\lfloor \frac{e_i - c_e - c_r}{c_e} \right\rfloor$ otherwise).

This leads us to a formal definition of $f_{M,\tau}$:

Definition 5. *The period allocation function $f_{M,\tau}$ used for a sensor just after it sent a message is defined by:*

- *if first message received from that sensor,*
  $$f_{M,\tau}(H_t) = \begin{cases} \tau|\Pi(t)| - (t - t_0)\%\tau & \text{if } |\Pi(t)| \leq M \\ \textit{death-date[0]} - t + M\tau & \text{if } |\Pi(t)| > M \end{cases} \tag{4.4}$$
- *Else,* $f_{M,\tau}(H_t) = \min(M, |\Pi(t)|)\tau$

The death-date list is updated to always contain the sorted list of sensor death instants whose relays are not covered. In particular, when a sensor comes alive while $|\Pi(t)| \geq M$, the death date of the sensor whose relay has just been taken is replaced by the predicted death of the new sensor. One property of this list is that its size never exceeds M.

Note that for the special cases $M = 1$ and $M = +\infty$ (all present sensors in round-

Algorithm 1 Death update algorithm

i is the ID of the transmitting sensor, e_i its energy just after it has transmitted the message and p_i its period which has just been set by $f_{M,\tau}$. The function called "add" (and "update") adds (and updates) elements to the list while sorting it in ascending date order.

Require: death-date, time t, sensor index i, period p_i, remaining energy e_i

if $|\Pi(t)| \leq M$ **then**
  if $e_i < c_e$ **then**
    Remove sensor i from death-date {i dies without another sensor taking over, the other sensors transmit in turn.}
  else if $p_i = M\tau$ **then**
    Update death-date of sensor i with value $t + p_i + M\tau \left\lfloor \frac{e_i - c_e}{c_e} \right\rfloor$
  else
    Update death-date of sensor i with value $t + p_i + M\tau \left\lfloor \frac{e_i - c_e - c_r}{c_e} \right\rfloor$
  end if
else
  if first activation from i **then**
    add death-date of sensor i with value death-date[0] $+ M\tau(1 + \left\lfloor \frac{e_i - c_e - c_r}{c_e} \right\rfloor)$ {i takes over after death-date[0], its death date is updated as if it were activating at a constant period of $M\tau$ afterward.}
    remove death-date[0] from death-date {The relay from death-date[0] is taken, so it is removed from the list.}
  end if
end if

robin), combinatorial and memory space simplifications can be done to implement $f_{M,\tau}$.

2.4 Properties of $f_{M,\tau}$

The following propositions establish that $f_{M,\tau}$ behaves as we wanted it to, with depiction of bounds on the sample span.

Proposition 2. *The $f_{M,\tau}$ period allocation function is effective on the instants of period τ.*

Proposition 3. *By simplifying the expression (removing the floor terms), and considering sensors with the same initial energy e, the sample span of $f_{M,\tau}$ is at least:*

$$L_{\min}(f_{M,\tau}) := \frac{ne - nc_e - (2n - 1 + M(M-1))c_r}{c_e} \tag{4.5}$$

and at most:

$$L_{\max}(f_{M,\tau}) := \frac{ne - nc_e - (2n - \mathbb{1}_{M=1})c_r}{c_e} \tag{4.6}$$

Proof (sketch). As long as there are no more than M alive sensors, in the worst scenario each new sensor that comes alive disrupts the existing schedule, forcing all other sensors to consume energy to change their activation period.

To get the upper bound, we on the contrary consider the most favorable scenario, that is when the first M sensors come alive in the same time interval of length τ, and n is a multiple of M. □

Formal demonstrations of Propositions 2 and 3 are respectively developed in Appendices B.2 and B.3.

The solution in the optimistic scenario is close to the global optimum $L_{\max}$ of Proposition 1.

In general, for a fixed τ, increasing the parameter M results in a higher number of period changes. In the worst case, this increase follows a quadratic relationship with respect to M. Since period changes are accounted for in our sensor energy consumption model, a higher number of period changes leads to increased sensor energy consumption, consequently reducing the overall monitoring duration. Additionally, though not measured here, it is worth noting that downlink transmissions need to be limited in constrained networks. For instance, in LoRa networks, adherence to duty cycles on the gateway side can

constrain the number of commands given to sensors, and antennas operating in downlink mode cannot listen to the uplink, potentially significantly impacting the overall Quality of Service [DHT19].

3 Simulations

This section discusses the experimental analysis of the period allocation function $f_{M,\tau}$, carried out through simulations. We propose to study the performance by using the function $f_{M,\tau}$ for different values of the number M of sensors jointly activating and of the target activation period τ. From the initial conditions defined in Table 4.1, we apply the period allocation function $f_{M,\tau}$ for each activation of sensor until the end of the monitoring, in order to determine monitoring duration and average diversity performance indicators.

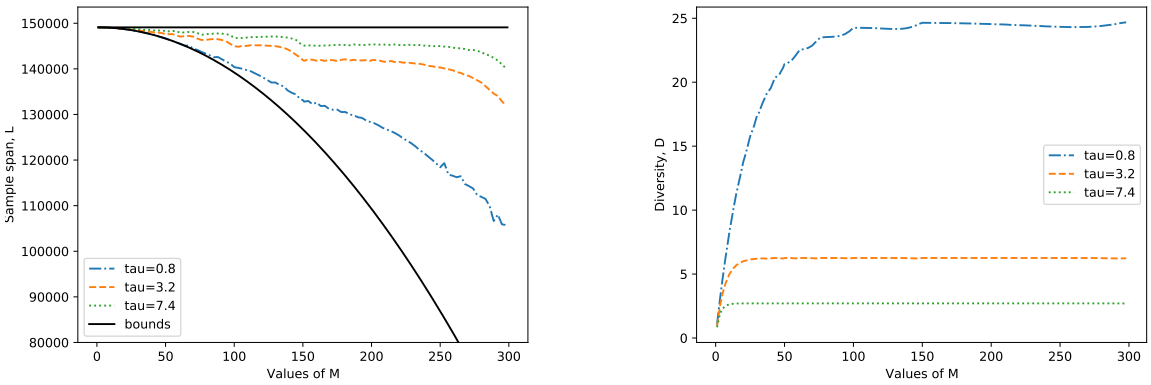
Parameter	Meaning	Value
n	Number of sensors	300
$e_i = e$	Battery capacity	500
$c_e = c_r$	activation and reception energy cost	1
$t_i - t_{i-1}$	Time between 2 consecutive initializations	$15 * 3.14(\pi)$
T	Relevance time of a data	20
Freshness function	Depletion of the data over time	v_{20}

Table 4.1: Simulation parameters

3.1 Influence of the Number M of Sensors Jointly Activating

For a fixed target activation period τ, the parameter M influences both performance metrics, as Fig. 4.3 illustrates. The sample span (thus, the monitoring duration) decreases when M increases (Fig. 4.3(a)), confirming the trends of the bounds developed in Proposition 3. Between $M = 1$ and $M = 300$, we observe a relative decrease of 6.2% of the total monitoring time for $\tau = 7.4$. It drops to 34.02% of relative difference for $\tau = 0.8$, since the sensors get scheduled more quickly and therefore are disturbed more times when new sensors come alive.

At a fixed τ, larger values of M offer greater diversity, as more sensors update their value periodically (Fig. 4.3(b)), with diminishing diversity gains as M increases (hence a concave function).



(a) (b)

Figure 4.3: Representation of some performance indicators using the period allocation function $f_{M,\tau}$, varying the number of jointly activating sensors M, for a few target activation periods τ. (a) corresponds to the sample span with analytical bounds, (b) the diversity. Each curve show values obtained over 100 simulation runs.

3.2 "Diversity Versus Duration" Trade-Offs

We represent in Fig. 4.4 the monitoring time and average diversity metrics obtained with $f_{M,\tau}$, for different values of (M, τ). Of course, one would like to be as north-east as possible in the figure (high diversity and high monitoring duration). Interestingly, the Pareto front is not always attained with the same value of M: if the network designer preferences (or the application needs) favor the monitoring duration, smaller values of M should be preferred, while larger values should be chosen if diversity matters most.

If the need for diversity is not very strong, then choosing a small value of M and a relatively large τ target activation period (compared to the relevance time of data T) allows to extend considerably the total monitoring duration (and induce a low consumption of the downlink). On the other hand, if the need for diversity is more important, it is necessary to choose a larger value of M, and a small target activation period τ, leading to a more frequent energy consumption, at the price of a shorter monitoring duration. As an example, if the diversity requirement is $D > 10$, then choosing $M = 44$ and $\tau = 1.97$, ensures the best monitoring duration, with $2.9 * 10^5$.

Hence the methodology leading to Fig. 4.4 can be adapted to the specific parameters of a new scenario, and applied to determine the best-performing parameters M and τ for the needs of the application.

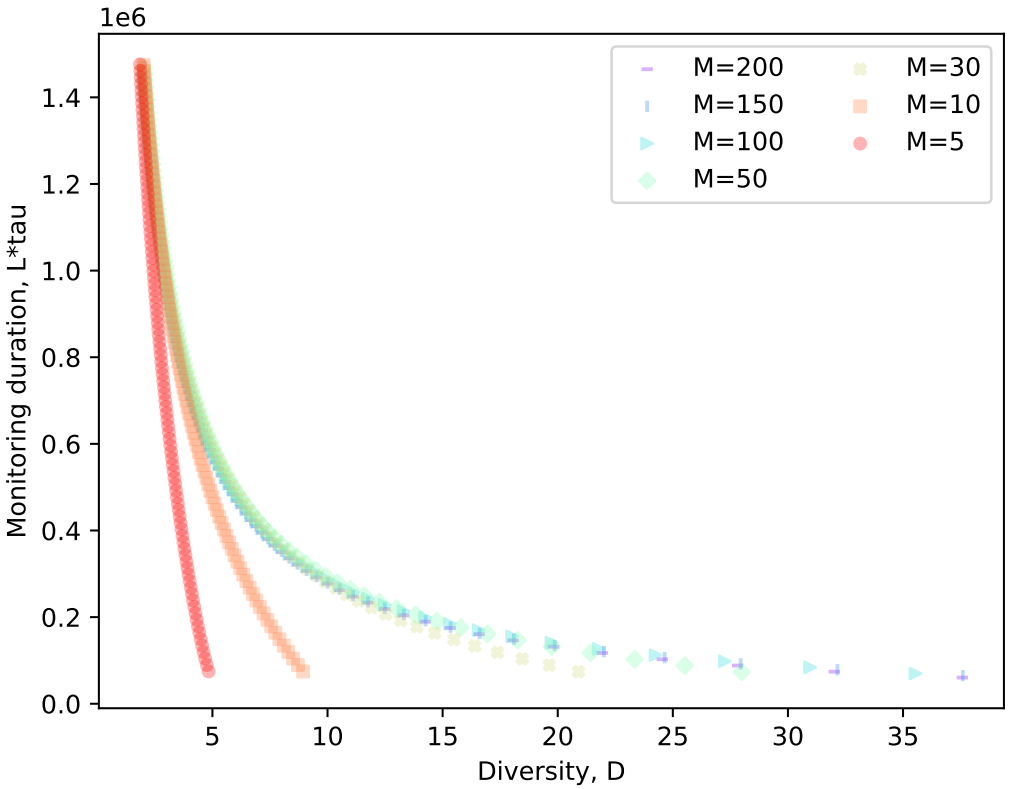


Figure 4.4: Performance of the update period function $f_{M,\tau}$, for several values of M (given in the legend) and τ (from 0.5 to 10 by 0.1 increments). Each point corresponds to the two-dimensional performance metrics (Diversity on the x-axis and monitoring duration on the y-axis), for fixed parameters M, τ of the update function.

4 Conclusion

This chapter lays the groundwork for the development of period allocation functions. We have introduced a function that dynamically manages the arrival of new sensors. It is composed of two key parameters: the time between message receptions and the number of sensors involved in the transmissions. Through simulations, we have established that these parameters need to be carefully chosen to strike a trade-off between energy efficiency and monitoring quality.

Our research findings have been published in [Mau+22].

Chapter 5

Asynchronous 2-Level Round-Robin Activations

In the previous chapter, we demonstrated that in a scenario where sensor observations are similar and frequent updates are desired, a trade-off must be made between tracking quality and system lifetime. Nevertheless, that solution is based on a strong assumption: the perfect knowledge of sensors' energy consumption, which allows predicting the instants when sensors deplete their battery. In reality, battery consumption is not so deterministic, and other independent factors can make a sensor out.

In this new section, we continue our research on period allocation functions. We consider the scenario where sensors can enter and exit the environment without prior notice. Building on the work presented in the previous chapter, we can only utilize the method that activates all sensors in a round-robin fashion, that defines our baseline. We mathematicaly prove that assigning the same activation period to all sensors maximizes average diversity, and having strictly periodic receptions maximizes minimum diversity. However, this solution has limitations, especially about the excessive number of period changes when dealing with a large number of changes in the sensor field. Additionally, the proposed synchronized scheduling approach (strict regular receptions) is challenging to implement in practice.

To address these issues, we develop a method that ensures overall regular sensor observations over time, at a user-selected rate, while minimizing management costs associated with sensor arrivals and departures. This solution is based on a binary tree structure for defining the allocation function.

By modeling arrivals and departures as random processes, we derive analytical approximations for the diversity metric based on the message reception rate parameter. Furthermore, by comparing our solution to the baseline function, which exhibits remarkable properties in terms of average diversity and limitation of minimum diversity, we demonstrate that relaxing the assumptions does not significantly affect the diversity met-

ric in practice. Overall, since this new function significantly reduces the number of period changes, it is better suited for the Massive IoT context considered in this research.

1 Problem Statement and Model

In this chapter, we consider a scenario that is largely similar to the one presented in Chapter 4. The similar assumptions are listed as follows:

- Sensors are deployed in the environment, and their initialization occurs during the monitoring process.
- Sensors have limited energy, which is consumed during activations.
- Sensors activate periodically, and their activation period can be updated following a sent message. We aim to define a period allocation function that can redefine a sensor's activation period after it sends a message.
- The freshness of a sensor is defined by the freshness of its last message, and diversity is a function defined over time as the sum of the freshness values of all sensors. The objective is to maximize this diversity.

Here, we want to propose a more realistic modeling approach for the initial condition of a sensor. In the previous chapter, we assumed that we knew the sensor's energy level and transmission consumption, allowing us to anticipate a sensor's death and manage it by taking over its tasks.

However, in this chapter, we consider that we don't know the sensor's energy level, hence we cannot anticipate a sensor's death. In our simulation model, we will assume that a sensor can die for two reasons. Firstly, it may deplete its battery, and transmitting more information increases the likelihood of its battery running out. The second reason is unrelated to the number of activations made: the sensor physically leaves the environment or experiences a hardware issue preventing it from providing relevant observations.

We no longer consider a sensor's period change in terms of its energy consumption; instead, we introduce it as a metric by tallying the total number of period changes. When a sensor needs to listen for an order from the terminal, it must keep its radio on for a longer time and employ signal processing algorithms to interpret the message, resulting in additional energy consumption. Moreover, the number of period changes

affects the terminal, as it cannot receive messages from sensors while transmitting an order [JR22]. This can significantly impact the quantity of messages received by the terminal. Furthermore, particularly within constrained networks, duty-cycle constraints may be enforced to regulate node activity, thereby restricting gateway communication to, for instance, only 1% of the time [DHT19].

In our simulation model, we assume that sensors continuously arrive in the environment, and we evaluate a solution over a specified duration. A period allocation function is thus evaluated based on two metrics, aiming to find a desirable trade-off. The objective for the user is to strike a balance by maximizing monitoring accuracy (represented by high diversity) while controlling management costs (the number of period change orders).

2 The Synchronized Round-robin Allocation Function

2.1 Definition of the Round-robin Function

Based on the work proposed in Chapter 4, we can extract a feasible solution under the more realistic assumptions of this new chapter. This baseline solution will serve as a reference for comparison with our novel contribution. Notably, this method exhibits remarkable properties in terms of maximizing average diversity.

Under the new assumptions of this chapter, it is no longer possible to evaluate the depletion of sensors. The only viable solution for $f_{M,\tau}$ occurs when $M = +\infty$, meaning the sensors are scheduled in a pure round-robin fashion. The period update function, denoted by $f_{\text{SRR},\tau}$ is:

$$\begin{aligned} \bullet \quad f_{\text{SRR},\tau}(H_t) &= \tau|\Pi(t)| - (t - t_0)\%\tau \quad \text{, if first message received from that sensor} \\ \bullet \quad &= \tau|\Pi(t)| \qquad\qquad\qquad\quad \text{, else} \end{aligned} \tag{5.1}$$

where H_t denotes the information gathered up to the current time. In this case, we store the value of $|\Pi(t)|$, which represents the number of present sensors, including the new sensors if any, and t_0 corresponds to the initialization time of the first sensor.

2.2 Properties

Although its expression is extremely simple, the function $f_{\mathrm{SRR},\tau}$ exhibits remarkable properties. Two key characteristics of this function are that (i) all sensors have the same activation period, and (ii) messages are received by the terminal at strict regular intervals.

Firstly, in order to maximize average diversity, all sensors must have the same activation period.

Proposition 4. *Assuming a freshness function that is strictly monotonic and differentiable, for a constant global reception frequency* $\sum_{i\in\Pi}\frac{1}{p_i}=\frac{1}{\tau}$*, all sensors must have the same period* $p_i=\tau|\Pi|$ *to maximize the average diversity.*

The proof of this proposition is based on the method of Lagrange multipliers and is developed in Appendix C.1.

On the other hand, strict regular reception allows for locally maximizing the minimum diversity.

Proposition 5. *For* n *sensors transmitting with a period of* $n\tau$*, at intervals of* τ*, shifting the activation time of a sensor away from its* τ *interval will decrease the minimum diversity.*

The proof of this proposition is based on the direct calculation of diversity, noting that the local minimums of diversity occur before each new activation. The proof is developed in Appendix C.2.

To summarize the properties of the Synchronous Round-Robin function:

- For n sensors and a fixed total number of messages per unit of time at $\frac{1}{\tau}$, transmitting with a period of $n\tau$ maximizes the average diversity and is the same regardless of the scheduling.

- If we consider a perfectly slotted scheduling, where sensors transmit in a pure round-robin fashion and the overall message transmission occurs at τ intervals, then it is the best local solution in terms of minimum diversity. Shifting the activation time of any sensor reduces the minimum diversity. Note that intuitively, we expect that this regular scheduling among sensors is also a global maximum for the minimum diversity over time, but we did not manage to prove it analytically.

3 The 2-Level Round-Robin Allocation Function

This section presents the contribution of the chapter, a period allocation function that maintains a constant overall period τ of data receptions with low number of period changes. With the function we propose, all sensors activate with a similar (up to a factor 2) activation period, while the terminal needs to send only one or two period change orders for each sensor arrival or departure.

Although this solution is no longer optimal in terms of average diversity since the periods are not the same for all sensors, this choice significantly reduces the number of period changes.

Furthermore, strict message reception synchronization is relaxed, resulting in a more varying diversity over time. The strict periodic reception of observations comes with practical costs: on one hand, it necessitates that a sensor receives two consecutive orders from the gateway to be included (which is not always guaranteed in practice), and on the other hand, sensors are susceptible to clock deviations that slightly shift their activation periods, thereby undermining strict scheduling. Similar observations have been made to critique synchronous routing solutions in [THH02], where all sensors turning on at the same time to relay information is shown difficult to achieve in practice.

3.1 Functioning Principle of the Allocation Method

To introduce the function, let us provide the global principle of functioning. We begin with a set of sensors that adhere to the global message reception property at the interval τ. The goal is to include and exclude sensors while minimizing the number of period change orders, all the while maintaining the property of receiving messages at the global average interval of τ.

When a new sensor is added, one of the existing sensors shares its message transmission load with the newcomer by doubling its activation period. The newly added sensor adopts this same new activation period.

On the other hand, when a sensor dies, the message emission load of the leaving sensor is added to the load of one of the sensors in use: its new activation period is defined so that the sensor activates as many times as both itself and the sensor that has just died.

3.2 Construction of the Period Allocation Function

To formalize this approach, we represent sensors as the leaves of a binary tree. The depth of a sensor in the tree represents its activation period in the form of $2^{\text{depth}}\tau$. Figure 5.1 illustrates how the tree evolves as sensors enter and leave the system.

The tree has a fundamental property: it is *full*, meaning that a node has either 0 or 2 children.

Furthermore, we impose that this tree is *complete*: the difference in depth between leaf nodes does not exceed 1. This limits the difference between activation periods by a factor of 2. Maintaining this property incurs a small additional cost in terms of period changes, which we elaborate on here.

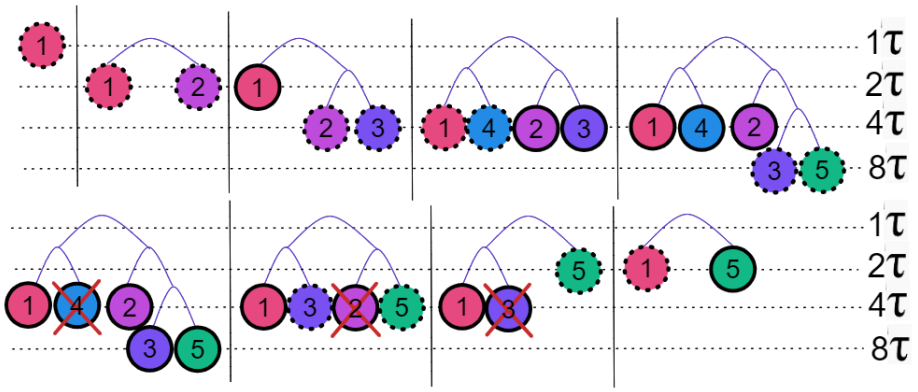


Figure 5.1: Evolution of the binary tree representation as sensors enter *(top)* or leave *(bottom)* the system; each sensor is represented with a colored circle with an ID, and horizontal dotted lines represent the activation periods of the sensors at that depth. A dotted line around a sensor means that its position (and height) was changed in the tree (hence a period change order is needed).
The top part represents the successive arrivals of sensors indexed from 1 to 5; the bottom one shows the successive departure of sensors 4, 2, and 3 (departures are symbolized by a cross).

Since the tree is balanced, the active sensors can therefore be grouped into two categories, whether their representation in the tree is in the last or second-last level in the tree, that we define as respectively **high-depth** and **low-depth** categories. We should note that if all the sensors belong to the same category, by convention we consider them all high-depth. Here we present the evolution of the representation tree when a modification occurs, such as the entry or exit of a sensor from the environment. This tree representation is then directly used for the definition of the period allocation function.

• *When a new sensor arrives in the environment*, if all sensors are high-depths, we now consider all of them being low-depths, so that in all cases there is a low-depth sensor. Hence, one of these (low-depth) sensors changes position by increasing its depth by one, and the new sensor becomes its *sibling* (with a same parent), both being high-depth thereafter.

• *When a sensor leaves*, there are 2 cases:

· If the leaving sensor is of high-depth, by construction it has a sibling, which becomes of low-depth by decreasing its depth by one. This is the case for the exits of sensors 2 and 3 in Fig. 5.1.

· If the leaving sensor is of low-depth, it is substituted with a high-depth sensor, whose displacement is treated like the departure of a high-depth sensor (described above). This is the case for the departure of sensor 4, replaced by 3 in Fig. 5.1.

For a sensor i that just sent a message, the period allocation function f that we suggest is then simply:

$$f_{\text{2LRR}}(H_t) = 2^{d_i}\tau \tag{5.2}$$

Where H_t represent the information collected up to time t. Here, we store and update the depth d_i of each sensor i.

3.3 Properties

We show here that our suggested period allocation function meets the objectives initially set, regarding the reception at a global rate of τ, with a limited number of period change orders over time. To that goal, we make the approximation that the period of a sensor of depth d is exactly $2^d\tau$ at any moment, while in reality, when a sensor changes positions in the tree (because of another sensor's arrival or departure), its activation period is only modified after its next activation.

For n sensors, let us denote by h the minimum depth of the tree, $h = \lfloor \log_2(n) \rfloor$. Then, according to the binary tree representation, we can say that:

• $n_{\min} = 2^{h+1} - n$ sensors activate at period $2^h\tau$ and are of low-depth. To understand this, if $n_{\min}$ additional sensors are added in the environment, they become complementary to each of the sensors of low-depth in order to make the binary tree *perfect*, with exactly $n + n_{\min} = 2^{h+1}$ sensors.

• $n_{\max} = 2(n-2^h)$ sensors activate with an activation period of 2^{h+1} and are of high-depth.

For instance, $n_{\min} + n_{\max} = n$.

Proposition 6. *At any moment, the average time between two sensor activations is τ. Mathematically, if Π denotes the current set of sensors in the tree, containing n sensors, and p_i the activation period of sensor i, we have:*

$$\sum_{i\in\Pi}\frac{1}{p_i}=\frac{n_{\max}}{2^{h+1}\tau}+\frac{n_{\min}}{2^{h}\tau}=\frac{2n-2^{h+1}}{2^{h+1}\tau}+\frac{2^{h+1}-n}{2^{h}\tau}=\frac{1}{\tau}.$$

Changing the position of a sensor in the tree results in a change of its activation period, ordered at its next activation. If the position is changed several times before a new activation, the sensor changes its activation period only once. Therefore, counting the number of position changes in the tree of a sensor provides us with an upper bound for the actual number of period change orders over time, a useful insight on the management cost of our method. From our tree construction, those position changes are quantified below.

Proposition 7. *When a new sensor arrives, the number of position changes in the tree (counting the position definition of the incoming sensor) is $r = 2$.*

When a sensor leaves the environment, the number of position changes is $r = 1$ if the sensor that dies is of high-depth and $r = 2$ if it is of low-depth.

4 A Markovian Model for Performance Evaluation

In this section, we develop a model to analyze as a Markov chain the evolution over time of the number n of jointly used sensors. This will be used to estimate the steady-state values of our performance metrics for the 2-level round-robin period allocation function.

4.1 Modeling Sensor Arrivals and Departures

We model sensor arrivals as a Poisson process, with an average arrival rate of λ sensors per time unit.

Regarding departures, we assume that a sensor can leave the environment for two main reasons:

• The sensor has consumed all its energy and switches off. We consider that the sensor has an initial energy which follows an exponential law with mean c_e/γ, with c_e the energy consumed for each activation, and γ a parameter characterizing the variability of the

battery state when joining the environment. To have a continuous-time Markov chain, we slightly relax the periodic-activation assumption from sensors, by assuming that each sensor i with period p_i transmit messages according to a Poisson process with rate $1/p_i$ (note that in our simulations, activations are really periodical). With this model, the time before running out of battery follows an exponential law of parameter $\frac{\gamma}{p_i}$. At any moment, the time before one sensor leaves because of battery depletion is then exponentially distributed, with parameter $\sum_{i \in S} \frac{\gamma}{p_i} = \gamma/\tau$, thanks to Proposition 6.

• The sensor leaves the environment because it has been physically removed, turned off, or has undergone a technical failure. For each sensor, the time before this occurs is modeled through an exponential law of parameter μ, hence with n sensors the time before one departure for this reason is exponentially distributed with parameter $n\mu$.

With those assumptions, the continuous-time process describing the number n of sensors in the system is a Markov chain, whose transition diagram is displayed in Fig. 5.2.

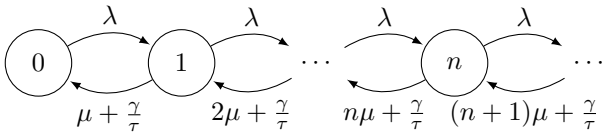


Figure 5.2: Continuous Markov modeling of the number of active sensors over time

4.2 Performance Metrics Estimation For 2LRR

We now use the Markov chain previously described to derive the steady-state distribution on n, and corresponding expected values for our performance metrics (approximating the actual ones).

4.2.1 Number of Sensors in the Steady State

Denoting by π_n the steady-state probability of having n active sensors, we have $\lambda\pi_{n-1} = (n\mu + \frac{\gamma}{\tau})\pi_n$ for all $n \geq 1$, leading to:

$$\pi_0 = \frac{1}{1+\sum_{n=1}^{+\infty}\left(\prod_{j=1}^{n} \frac{\lambda}{j\mu+\frac{\gamma}{\tau}}\right)}$$
$$n \geq 1, \quad \pi_n = \left(\prod_{j=1}^{n} \frac{\lambda}{j\mu+\frac{\gamma}{\tau}}\right)\pi_0$$

4.2.2 Number of Period Changes

The number $\dot{r}_n$ of position changes of sensors in the tree per time unit if there are n sensors is, by splitting between low and high-depth sensors (from Proposition 7):

$$\dot{r}_n = 2\left(\frac{\gamma}{\tau}\frac{2n_{\min}}{2n_{\min}+n_{\max}} + n_{\min}\mu\right) + \left(\frac{\gamma}{\tau}\frac{n_{\max}}{2n_{\min}+n_{\max}} + n_{\max}\mu\right) + 2\lambda$$

An upper bound for the average number of period change orders sent per time unit is then:

$$\dot{r} = \sum_{n=1}^{+\infty} \pi_n \dot{r}_n$$

4.2.3 Mean Diversity

Considering the freshness function $u_T(x) = e^{-\frac{x}{T}}$, the average diversity for one sensor of activation period p is:

$$\frac{1}{p}\int_0^p e^{-\frac{t}{T}}dt = T/p(1-e^{-p/T})$$

Then, we can estimate the average diversity D_n for n sensors as

$$D_n = Tn_{\max}\frac{1-e^{\frac{-2^{h+1}\tau}{T}}}{2^{h+1}\tau} + Tn_{\min}\frac{1-e^{\frac{-2^{h}\tau}{T}}}{2^{h}\tau},$$

and the (steady-state) average diversity D as

$$D = \sum_{n=1}^{+\infty} \pi_n D_n. \tag{5.3}$$

5 Simulation Results

This section compares the 2-level round-robin method developed in this chapter to other strategies, highlighting that it is the best fitted method under the hypotheses and objectives considered. Moreover, we show that the analytical study can help find the user parameter τ maximizing the diversity.

Parameter	Meaning	Value
	Start of diversity acquisition	$20000s$
	Duration of the first phase	$50000s$
	Duration of the second phase	$50000s$
λ_1	Sensor arrival rate - first phase	$0.1s^{-1}$
λ_2	Sensor arrival rate - second phase	$0.001s^{-1}$
$1/\gamma$	Average number of send messages	1000 activations
μ	Rate of departure for other reasons	$0.00001s^{-1}$
Freshness	Value depletion with time x	$e^{-x/T}$
T	Relevance time of data	$100s$

Table 5.1: Simulation parameters

5.1 Comparative Performance Evaluation

5.1.1 Simulation Setting

We consider an initially empty system, with sensors entering and leaving as per the random processes described in Section 4.1, except activations are really periodic. We assume two consecutive phases: in the first one, many sensors enter the environment, while in the second one, sensors enter the environment more rarely. We start observing the environment (i.e., computing the metrics) after an initialization time. Through this simulation, our method should save the energy of the sensors when they are in high density, so as to ensure a better diversity (because more sensors will still be alive) when sensors enter more occasionally.

For all three methods, we apply the period allocation function after each sensor message reception, and evaluate the overall performance after the simulation is completed. The parameters of the simulation are given in Table 5.1.

Recall our two metrics are diversity (that varies over time) and management cost (overall number of period update orders). Rather than the average diversity value over time, we display here its 5th percentile, that is, the diversity value that is guaranteed 95% of the time. For the management cost, we just count the period update orders sent per time unit.

5.1.2 Other Scheduling Methods for Comparison

To compare our two-level round-robin period allocation function, we also implement the synchronized round-robin function. Both functions are parameterized by τ, which corre-

sponds to the (average) time between two sensor activations. The synchronized round-robin function has the property of maximizing the average diversity when aiming to receive a constant number of messages per unit of time.

Moreover, we also propose to implement the simplest sensor management method, that fixes the same (given) activation period p to all newly arrived sensors. This method can represent a non-slaved solution, where we do not modify the initial period of a sensor, although in our simulations, we will evaluate the quality of the solution for different choices of p. That method, that we call **static**, minimizes the number of period change order, but does not adapt to the changing number of present sensors.

5.1.3 Performance Evaluation

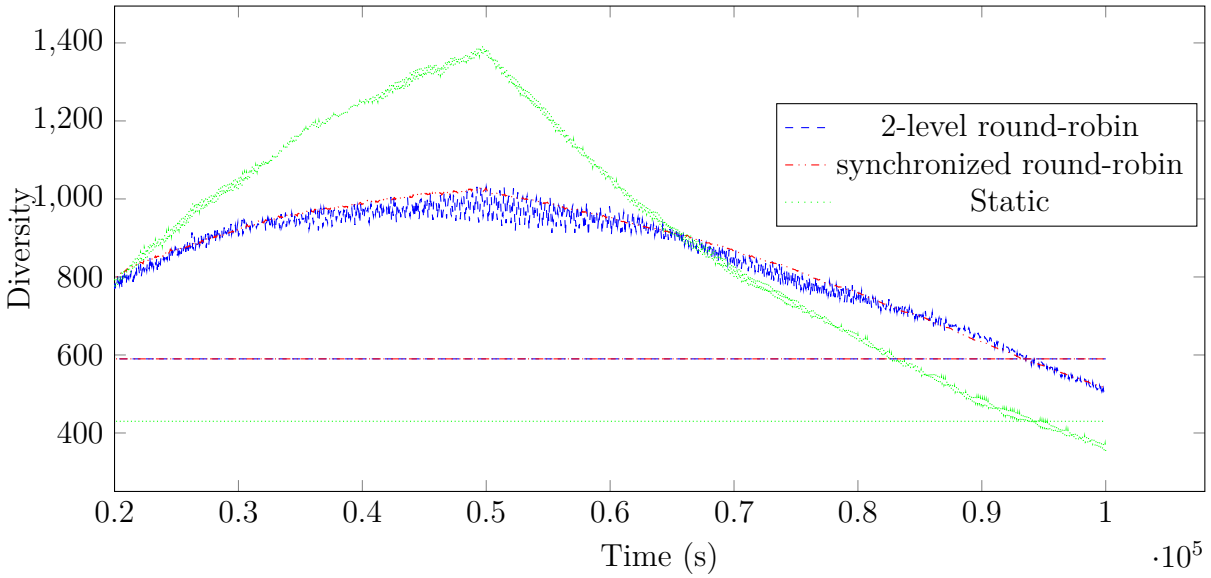


Figure 5.3: Diversity over time for a simulation, with global period $\tau = 0.1$s for round-robin methods and individual period $p = 150$s for the static method, and diversity guarantee (fifth percentile) for each method (horizontal lines).

Fig. 5.3 illustrates a simulation trajectory, showing the diversity over time for the 3 management methods, with parameters $\tau = 0.1s$ for the two round-robin methods, and $p = 150s$ for the static one. The curves show how the period allocation function manages sensor activations, in particular how it adapts to sensor field changes. We graphically show our overall diversity metric, that is the 5th-percentile over the observation period: 95% of the time, the instantaneous diversity exceeds that value.

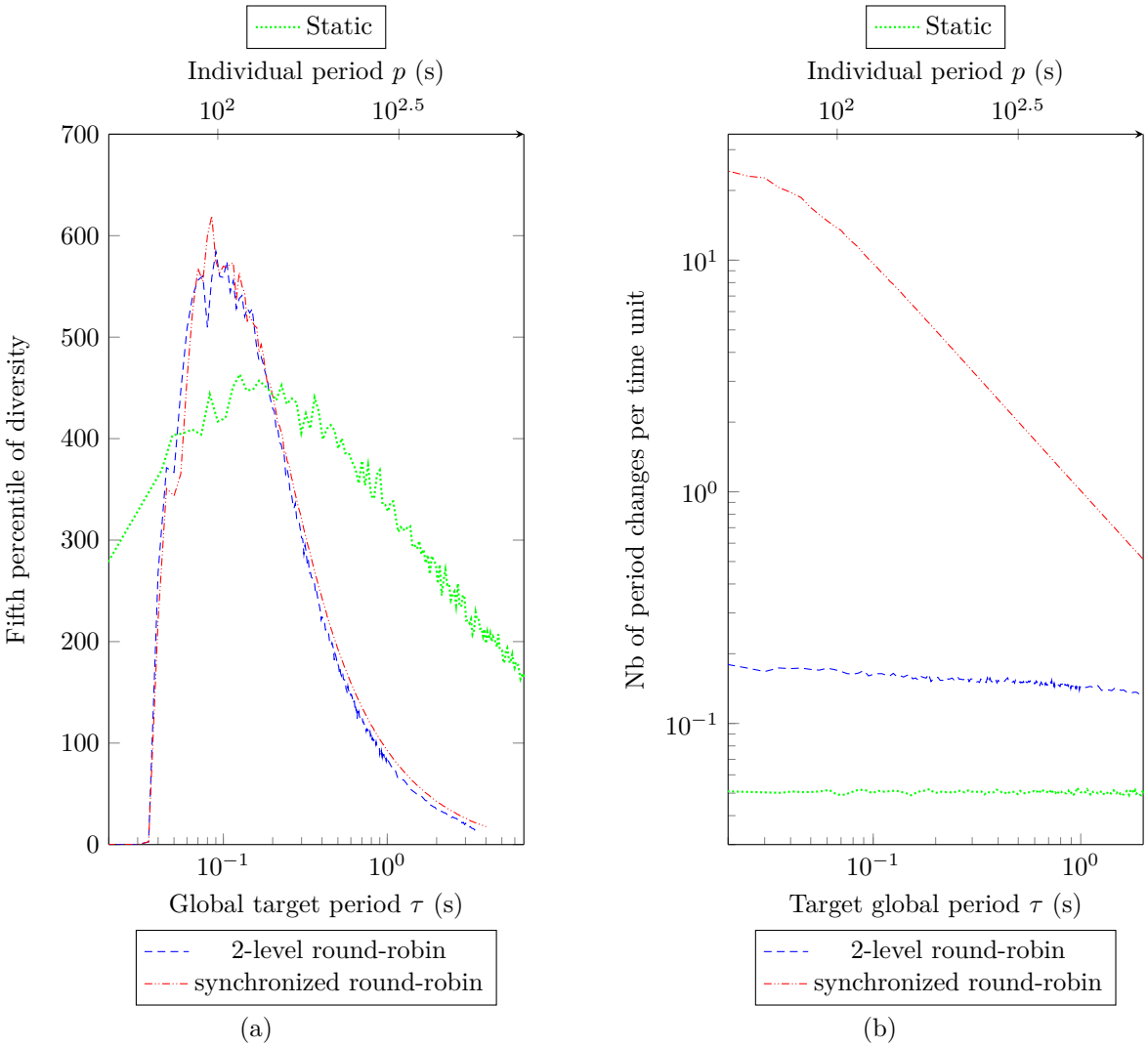


Figure 5.4: Fifth percentile of diversity (a) and number of period change per time unit (b) for the three methods, versus their parametrization. Parameters for round-robin methods are in bottom x-axis and in top x-axis for static.

The overall performance metrics of the three methods for different parameters are shown in Fig. 5.4(a,b), for different parameter values (τ on the bottom x-axis for round-robin methods, and p for static on the top).

From Fig. 5.4(a), the best methods under the simulation conditions are the round-robin ones, each insuring the best monitoring quality for τ around 0.1s, with minimum 95% diversity guarantee of 578. The static method performs less well in our simulations, even with the most favorable fixed period p: the maximum diversity guarantee is 420 with $p \simeq 150$, i.e. a 30% lower performance. One reason for this is that it does not adapt to the number of present sensors, hence may overuse the sensors when there is a high density, rather than saving their energy for later.

Note that although the synchronized round-robin should lead to a more stable and better diversity due to strict periodic message receptions and similar activation periods to all senors, both round-robin methods provide fairly similar diversity over time.

However, those strict periodic receptions come with a high management cost, as illustrated in Fig. 5.4(b). For $\tau = 0.1$, synchronized round-robin implies 59 times more period update messages to the sensors than 2-level round-robin. This is due to our tree structure, that limits the number of period change orders to 1 or 2 for each arrival or departure, instead of n for synchronized round-robin .

5.2 Search for the Optimal τ Parameter

We show here how to choose the parameter τ to have the best monitoring quality, in a steady-state situation (we take here the second phase of our simulation, as an example). In Fig. 5.5(a), we show the instantaneous diversity over time when $\tau = 5$, with a steady-state behavior around the theoretical expected value computed in (5.3) from the Markovian model. In Fig. 5.5(b), we compare that theoretical mean diversity from (5.3) with the simulated fifth percentile for different values of τ.

From these results, if sensor arrivals and departures are reasonably modeled with Markovian models, then we can approximate the mean diversity in the steady state, for a given user parameter τ. This can be used to choose a well-performing τ, which should also be close to optimal for the fifth percentile, as suggested by Fig. 5.5 (b).

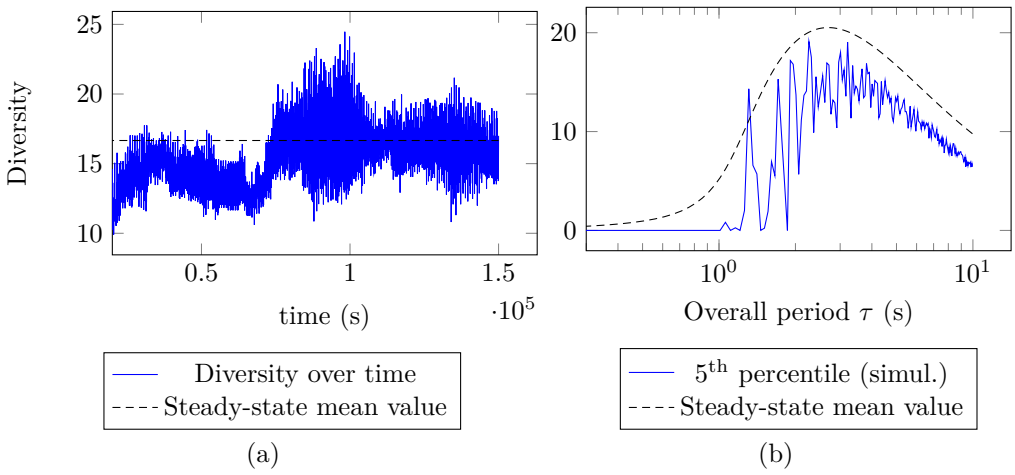


Figure 5.5: Simulation versus theoretical results, with constant parameters for sensor arrivals and departures. One simulation trajectory of diversity over time is shown in (a), with the corresponding steady-state mean value for the Markovian model, for $\tau = 5$. In (b) we compare the fifth percentile of simulated diversity with the theoretical mean diversity: both reach their maximum for approximately the same global period τ.

6 Conclusion and Discussions

In this chapter, we have introduced a novel period allocation function that enables the reception of messages at globally regular time intervals using the observation period as its sole parameter. This new approach significantly reduces the frequency of period changes, as compared to the approach presented in Chapter 4 (which emphasizes maximal accuracy properties) and offers greater adaptability by not mandating strictly periodic message receptions. Through simulations, we have demonstrated that relaxing the requirement for uniform sensor periods and strict message reception has minimal impact on our accuracy metric. By modeling sensor arrivals and departures as Poisson processes, we have extracted theoretical insights that provide approximations of average diversity. These insights aid in the pursuit of the optimal target observation period parameter that maximizes diversity.

The period allocation methods proposed in this thesis part have a specific objective: to complement similar sensor clustering solutions, aiming to distribute the observation load within clusters of identified similar sensors. Given the temporal variability of the sensor fleet and the emergence and disappearance of similarity links, these clusters are inherently sensitive to changes. Consequently, our focus has been on the robustness to such changes, ensuring effective load distribution within clusters of similar sensors. Notably, in the latest

version presented in Chapter 5, we have introduced mechanisms that significantly reduce the frequency of period changes when modifications occur within the sensor cluster (such as sensor additions or removals).

Our research findings have been published in [Mau+23].

Part III

SIMILARITY METRIC AND COVERING SUBSET ALGORITHM

Chapter 6

Grouping Sensors Based on Observations

In this part, encompassing a single chapter, we present a contribution to the similarity metric and covering subset algorithm components. Specifically, we propose an evaluation of sensor similarity based on their observed data, followed by the creation of clusters of similar sensors. Our focus is on a generic scenario that has not been extensively explored in existing literature: sensors are deployed in the environment for a finite duration and transmit noisy observations irregularly over time, without synchronization between them.

Firstly, we propose a distance metric that employs the Kriging interpolation method to evaluate the differences in average magnitude between interpolations. Additionally, we introduce a hierarchical clustering solution that groups similar sensors, proposing a linkage method that assigns higher weights to distances calculated over longer durations.

Through simulations, we demonstrate that our distance metric effectively distinguishes historical observations following the same phenomenon from those following different ones, outperforming the Jaccard index, which is our comparative metric due to its suitability for our assumptions. Furthermore, we adapt an existing solution from the literature, aiming to limit the maximum discrepancy between all interpolations within the same cluster. We establish the superiority of our approach in terms of the quality of sensor grouping.

1 Sensors, Observation Histories, and Objectives

In this section, we outline the objective and the hypotheses concerning the deployment of sensors and their observations.

1.1 Identifying Sensors Belonging to the Same Phenomenon

Sensors are deployed in a given environment, which can be divided into multiple distinct phenomena, each exhibiting specific characteristics. For instance, in a building, temperature variations may differ from one room to another.

Our objective is to group sensors that observe the same phenomenon. Initially, we aim to define a similarity metric to assess the proximity between sensors. Subsequently, we develop a clustering solution to group together sensors deemed similar. The broader goal (not done in the chapter) is to establish observation collection schemes that leverage similarity to reduce the amount of information transmitted by sensors, thereby conserving their energy resources.

1.2 Incoming and Outgoing Sensors

We consider scenarios where sensors enter and exit the environment dynamically. For example, new sensors might be added over time, while others may become inactive due to hardware issues or depleted batteries. For instance, in a logistics exchange platform, sensors may be associated with specific shipments and can stay only temporarily in the environment.

Consequently, the similarity between two sensors can only be evaluated when they are coexisting within the environment. Notably, this shared time interval of operation is variable or even non-existent.

1.3 Observations Sent by the Sensors

A sensor provides observations of a particular phenomenon, which it directly transmits to the terminal. Here, the sensors are transmitting noisy observations due to imprecise measuring devices.

It has been shown in [THH02], that activating sensors on the same instants is highly costly. Hence, we assume that the sensors cannot be synchronized to operate on jointly defined time steps.

Furthermore, we assume that sensors send observations irregularly. The specific data collection method employed by a sensor (e.g., trigger-based, model-based [DBO17], similarity-based) influences the observation period, which tends to vary over time.

1.4 Observation History Definition

An **observation history** refers to a set of observations generated by a sensor, each observation being composed of a time and a corresponding value. Our investigation centers on the transmissions conducted by these sensors; henceforth, moving forward, the object for which we intend to assess similarity and subsequently group are denoted as observation histories. We refrain from using the term "time series" because, although related, time series assume observations made at the same time instants and regular intervals, which is not the case in our context.

2 Defining a Distance Metric Based on Sensor Observations

As explored in Chapter 3 Section 3, there exist similarity metrics proposed based on sensor observations. However, none of these metrics are capable of fully satisfying the constraints inherent to an MIoT deployment, which include realistic environment, accommodating non-synchronized observations, managing noisy observations, and identifying compromised sensors. An example of two observation histories to be compared is depicted in Fig. 6.1

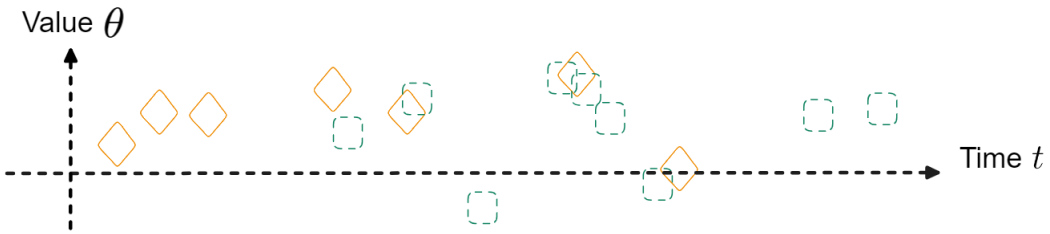


Figure 6.1: Representation of two observations histories.

In this section, we introduce a distance metric that relies on two key components. Firstly, we utilize an interpolation method to convert irregular observations into a continuous representation. Subsequently, we define the distance between two interpolations over their common time interval by the average magnitude difference.

2.1 Interpolation Function Based on an Observation History

The historical observations are irregularly spaced and noisy, making direct comparisons challenging. Therefore, as an initial step, we propose to employ an interpolation method on each observation history. This approach transforms the observations into continuous functions defined over intervals, facilitating comparisons.

Justification of the Kriging Choice

An interpolation function is a mathematical function defined over all time points based on a set of noisy observations. Its objective is to minimize the average discrepancy between the interpolated function and the measured phenomenon. Numerous interpolation methods exist, as documented in [CQ98].

Since the observed data is subject to noise, we aim to relax the constraint of passing through all data points. Consequently, certain methods like Spline are not applicable.

Kriging is an interpolation method based on Gaussian processes governed by prior covariances [Kle09]. This approach is particularly well-suited for various noise reduction applications, as summarized in [PWG13], as it allows the estimation and incorporation of measurement errors into the modeling. For instance, in [Zim+99], an experimental study demonstrated the superiority of Kriging over the inverse distance weighting method. Kriging has been applied in the domain of the IoT as well, such as in [Cas+10], where it was used to propose a sensor positioning solution based on the data they provide.

Principle of the Variogram

The variogram is a function based on a statistical model, where each observation is considered to be a random variable. It quantifies the correlation between two observation values according to their temporal separation. It is used in the kriging model to evaluate the correlation between a value to interpolate for a target time and the known observations.

Since the true variogram is typically unknown, it is estimated using known observations. This estimation is obtained by initially calculating the experimental variogram. We denote by $\theta = \{\theta_t, t \in T\}$ an observation history, where T represents the set of measurement time instants and θ_t is an observation made at time t. Then, the experimental variogram γ is computed for each pair of points, so that:

$$\forall (t_1, t_2) \in T^2, \gamma(|t_1 - t_2|) = 0.5(\theta_{t_1} - \theta_{t_2})^2$$

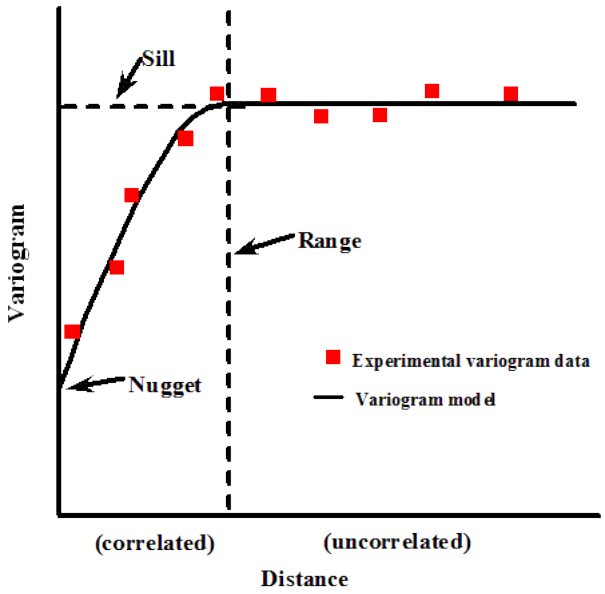


Figure 6.2: Diagram illustrating the variogram model based on experimental variogram points. The variogram consists of three parameters: nugget, sill, and range.

The points of the experimental variogram are depicted in Fig. 6.2. To obtain a continuous representation from this discrete experimental variogram, we fit these points to a function referred to as the **variogram model**, denoted by $\hat{\gamma}$. This model allows us to characterize the correlation between two observations based on their temporal separation.

For example, spherical, exponential, and Gaussian models are characterized by three parameters and illustrated in Fig. 6.2:

- The nugget n: Reflects the amount of short-range variability in the data. It is related to the measurement noise.
- The sill s: Represents the value at which the variogram levels off.
- The range r: Denotes the lag distance at which the variogram reaches the sill value.

The Gaussian variogram is given by:

$$\hat{\gamma}(\Delta_t) = n + s\left(1 - e^{-\frac{\Delta_t^2}{r^2}}\right) \tag{6.1}$$

Calculations for the Simple Kriging

Kriging is an interpolation method rooted in statistical modeling. It assumes that each observation is a random variable with a finite mean and variance.

Here, $\theta = (\theta_t)_{t \in T}$ constitutes the vector representing the observation history. The covariance matrix of the observation history vector is defined using the variogram $\hat{\gamma}$ as follows: $K = E[\theta\theta^\top] = (\hat{\gamma}(t_1, t_2))_{t_1, t_2 \in T}$.

Our objective is to evaluate the value at the point $\hat{t}$. Let $\Theta_{\hat{t}}$ denote the random variable representing the value at $\hat{t}$. The covariance vector between the value at $\hat{t}$ and each observation in the observation history is defined based on the variogram model: $k_{\hat{t}} = E[\theta\Theta_{\hat{t}}] = (\hat{\gamma}(\hat{t}, t))_{t \in T}$.

Based on these notations, we present the result for the simple kriging. The strong assumption here is that the mean expectation of values at all time instances is the same and known, assumed to be zero: $\forall t, E[\Theta_t] = 0$. In the case of ordinary kriging (another kriging modeling), the expectation is similar across all points and unknown; for universal kriging, a polynomial trend model is incorporated. By subtracting the general function, we get back to the basic case of solving simple kriging.

The core principle of kriging is that interpolation at a point is defined as a linear combination of the observed values. Hence, the estimator at the point $\hat{t}$, denoted by $\hat{\theta}_{\hat{t}}$, is the sum of observation values of the observation history, weighted by the coefficient vector $\psi_{\hat{t}} = (\psi_{t,\hat{t}})_{t \in T}$:

$$\hat{\theta}_{\hat{t}} = \sum_{t \in T} \psi_{t,\hat{t}} \theta_t = \psi_{\hat{t}}^\top \theta$$

The weights are defined to minimize the expectation of the squared difference between the estimator and the quantity to predict at this new point $\hat{t}$:

$$\Delta(\hat{t}) = E[(\hat{\theta}_{\hat{t}} - \Theta_{\hat{t}})^2] \tag{6.2}$$

Through the development of this equation, elaborated in Appendix D, we can deduce that:

$$\Delta(\hat{t}) = \psi_{\hat{t}}^\top K \psi_{\hat{t}} - 2\psi_{\hat{t}}^\top k_{\hat{t}} + \text{constant} \tag{6.3}$$

$\Delta(\hat{t})$ is minimized when $\psi_{t,\hat{t}} = K^{-1} k_{\hat{t}}$, leading to the estimation of the value at $\hat{t}$:

$$\hat{\theta}_{\hat{t}} = k_{\hat{t}}^\top K^{-1} \theta \tag{6.4}$$

The estimator at a given time $\hat{t}$ is the result of a matrix computation. Here, K represents the covariance matrix among the observations within the observation history, and $k_{\hat{t}}$ denotes the covariance between the observations from the observation history and the target value. All of these covariance values are computed using the variogram

model, which defines the correlation between observation values based on their temporal separation.

2.2 Distance Based on Mean Magnitude Difference

Let $i : \{\theta_{i,t}, t \in T_i\}$ and $j : \{\theta_{j,t}, t \in T_j\}$ be the observation histories under study, so that $\hat{\theta}_i(t)$ and $\hat{\theta}_j(t)$ be the interpolations obtained using Kriging. We use the mean magnitude difference to evaluate the distance between two interpolations, as schematically represented in Fig. 6.3.

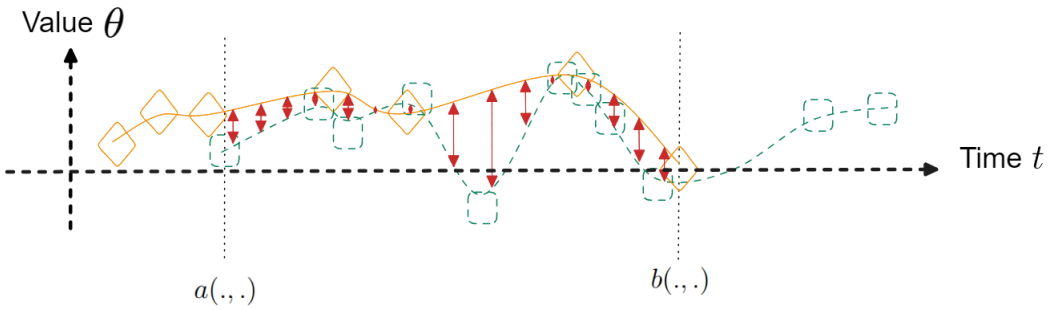


Figure 6.3: Orange diamonds and dashed green squares represent two observation histories, with time on the x-axis and observation values on the y-axis. The interpolations are depicted as solid orange and dashed green lines, respectively. The vertical dashed lines indicate the common temporal domain of the two interpolations $[a(.,.), b(.,.)]$. The distance between the two interpolations is defined by the mean distances (red arrows) over the common definition interval.

Firstly, the interpolations can only be compared over their common definition interval. If there exists a common definition interval between i and j, we denote it by $[a(i,j), b(i,j)]$. This interval begins at the time of the first observation in the observation history that started the latest, and it ends at the time of the last observation in the observation history that ended the earliest:

$$\begin{aligned} a(i,j) &= \max\{\min\{t \in T_i\}, \min\{t \in T_j\}\} \\ b(i,j) &= \min\{\max\{t \in T_i\}, \max\{t \in T_j\}\} \end{aligned} \tag{6.5}$$

If the definition intervals do not overlap, i.e. $a(i,j) > b(i,j)$, it is not possible to define a distance between these sensors. Hence, the duration of the common definition interval, denoted by $\delta(i,j)$, is defined by:

$$\delta(i,j) = \max\{0, (b(i,j) - a(i,j))\} \tag{6.6}$$

Furthermore, since the interpolation method aims to minimize the average difference between the ground truth and the estimation, we define the distance $d(i,j)$ as the mean magnitude difference between the interpolations. If the duration of the common definition interval is not zero, it can be mathematically expressed as:

$$d_{\text{interp-mean}}(i,j) = \frac{1}{\delta(i,j)} \int_{a(i,j)}^{b(i,j)} |\hat{\theta}_i(t) - \hat{\theta}_j(t)| dt \tag{6.7}$$

3 Tuned Linkage Hierarchical Clustering

In the state-of-the-art chapter (Chapter 3 Section 4), we concluded that for the covering subset algorithm component, it is preferable to use methods based on grouping sensors considered similar. In this section, we propose a method that relies on a continuous similarity measure to cluster together observation histories that are considered similar, using a hierarchical clustering approach.

3.1 Specification of the Clustering Problem

In a conventional clustering problem, we consider objects with n variables and aim to group together objects that are close when represented in a space where each variable constitutes a dimension [Déj+07; ASY15; Lia05].

In our context, an object represents an observation history and, consequently, an interpolation defined over a time interval. These objects can no longer be represented in a space of n dimensions. The distances between objects are not so straightforward, which is why we have dedicated a specific section to define the distance between two objects. For a pair of observation histories, we have defined a **distance** $d(.,.)$ on their **common definition interval duration** $\delta(.,.)$.

This change implies specific considerations in devising a clustering solution:

- Some objects may have an unknown distance: they are defined over disjoint intervals, making it impossible to determine their proximity,
- The comparison time frame is an essential indicator for defining the quality of the distance measure: a distance calculated over a longer definition duration carries

more significance than one computed over a very short time span.

3.2 The Agglomerative Hierarchical Clustering

Algorithm Principles

For this problem, we choose to focus on solutions based on agglomerative hierarchical clustering. This clustering method involves iteratively merging clusters together [LW67].

Initially, each object (observation history) is considered as its own cluster. At each iteration, the two closest clusters are merged to form a new cluster. Consequently, in each iteration, we obtain one less cluster than in the previous iteration. The stopping criterion for merging is either when a distance threshold is exceeded or when the desired number of clusters is reached.

The Linkage Method

An essential aspect here is the definition of the distance between clusters. The method that relies on inter-object distances to determine the inter-cluster distance is referred to as the **linkage method**. In Fig. 6.4, we illustrate several linkage methods: Simple-link defines the distance between clusters as the smallest distance between any pair of objects from a different cluster; complete-link uses the largest distance between any pair of objects from a different cluster; average-link calculates the average of all pairwise distances between objects from a different cluster.

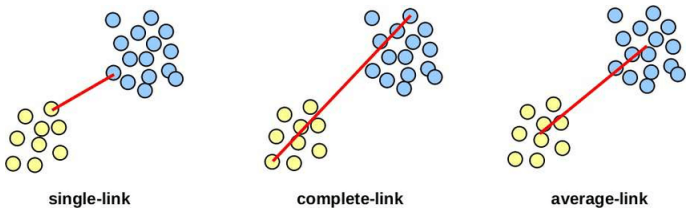


Figure 6.4: Examples of linkage methods, from [Gue11].

As shown in Fig. 6.4, the distance between clusters is not always defined by all distances between individual objects. Therefore, by tuning the linkage method, it is possible to achieve a clustering method that takes into account unknown distances between objects.

3.3 Tuning of the Linkage Method

In the literature, various common linkage methods exist, all of which involve linear combinations of distances between the observation histories within the clusters being compared.

Here, we choose to adapt the average-link to better suit our problem. We weigh the distances by the duration of the common definition interval to give more importance to distances calculated over longer periods.

Let $d(i,j)$ be the distance between observation history i and j calculated using the method described in Eq. (6.7), and $\delta(i,j)$ be the duration of their common definition interval, as defined in Eq. (6.6). When two observation histories are not directly comparable, $\delta(i,j) = 0$, and $d(i,j) = None$, and our convention dictates $\delta(i,j)d(i,j) = 0$ in order not to involve unknown distance in the linkage method.

We define the distance between two clusters as the sum of distances between pairs of objects from different clusters, weighted by their common definition interval duration. Considering $i \in I$ as the set of observation histories included in cluster I, and $j \in J$ for J, the distance between clusters I and J is given by:

$$D(I,J) = \frac{\sum_{i\in I}\sum_{j\in J}\delta(i,j)d(i,j)}{\sum_{i\in I}\sum_{j\in J}\delta(i,j)} \tag{6.8}$$

(If all distances between i and j are unknown, then by convention, we will have $D(I,J) = None$, and we will not merge I and J.)

For this linkage method, we employed the Lance-Williams algorithm as a reference for hierarchical clustering construction [MC12]. This algorithm updates the distance between clusters at each merging step. Let us denote the cluster composed of elements from clusters I and J as $I+J$, and another cluster as K. Considering the total shared duration between all pairs of observation histories in clusters I and J, the update formulas are as follows:

$$\begin{aligned} D(I+J,K) &= \tfrac{\delta(I,K)}{\delta(I,K)+\delta(J,K)}D(I,K) + \tfrac{\delta(J,K)}{\delta(I,K)+\delta(J,K)}D(J,K) \\ \delta(I+J,K) &= \delta(I,K)+\delta(J,K) \end{aligned}$$

As a reminder of the agglomerative algorithm, in each round, we choose to merge clusters with the smallest distance D based on this distance definition.

4 Simulations

In this section, we perform simulations by generating two distinct continuous phenomena, each observation history consistently following one of the two phenomena. Specifically, an observation is the value of the corresponding phenomenon at the time of measurement, with added noise.

We model the characteristics of observation histories using exponential laws, such as creation of new observation histories and their duration, and the observations made over time. We vary the measurement noise to study the extent to which our solution can identify similarities and group observation histories following the same phenomenon.

Firstly, we demonstrate that our distance definition better identifies observation histories that follow the same phenomenon compared to a distance metric based on sets from the literature.

Next, we adapt a solution from the literature to our constraints, aiming to group observation histories while limiting the maximum differences between observations within the same group. We show that an approach based on means, as proposed in this chapter, is preferable.

4.1 Generation of Phenomena and Observation Histories

Generation of Phenomena

We define a phenomenon using a continuous function over time. In this study, we consider two phenomena, each generated in the same way. Specifically, the generic function is given by:

$$f(t) = \sum_{i=1}^{30} (\alpha_i \cos \omega_i t + \beta_i \sin \phi_i t)$$

For each $i \in \{1, 30\}$ and for each of the two phenomena, the constants α_i and β_i are chosen from a uniform distribution $\mathcal{U}(-100, 100)$, and the frequencies ω_i and ϕ_i are chosen from a uniform distribution $\mathcal{U}(0, \frac{2\pi}{30})$ (ensuring a minimum oscillation period of 30, limiting the variability). Then, we rescale the function to the range $[-1, 1]$ (compressing the phenomena values into a small value segment). We keep the same phenomena for all the simulation parts, and they are depicted in Fig. 6.5.

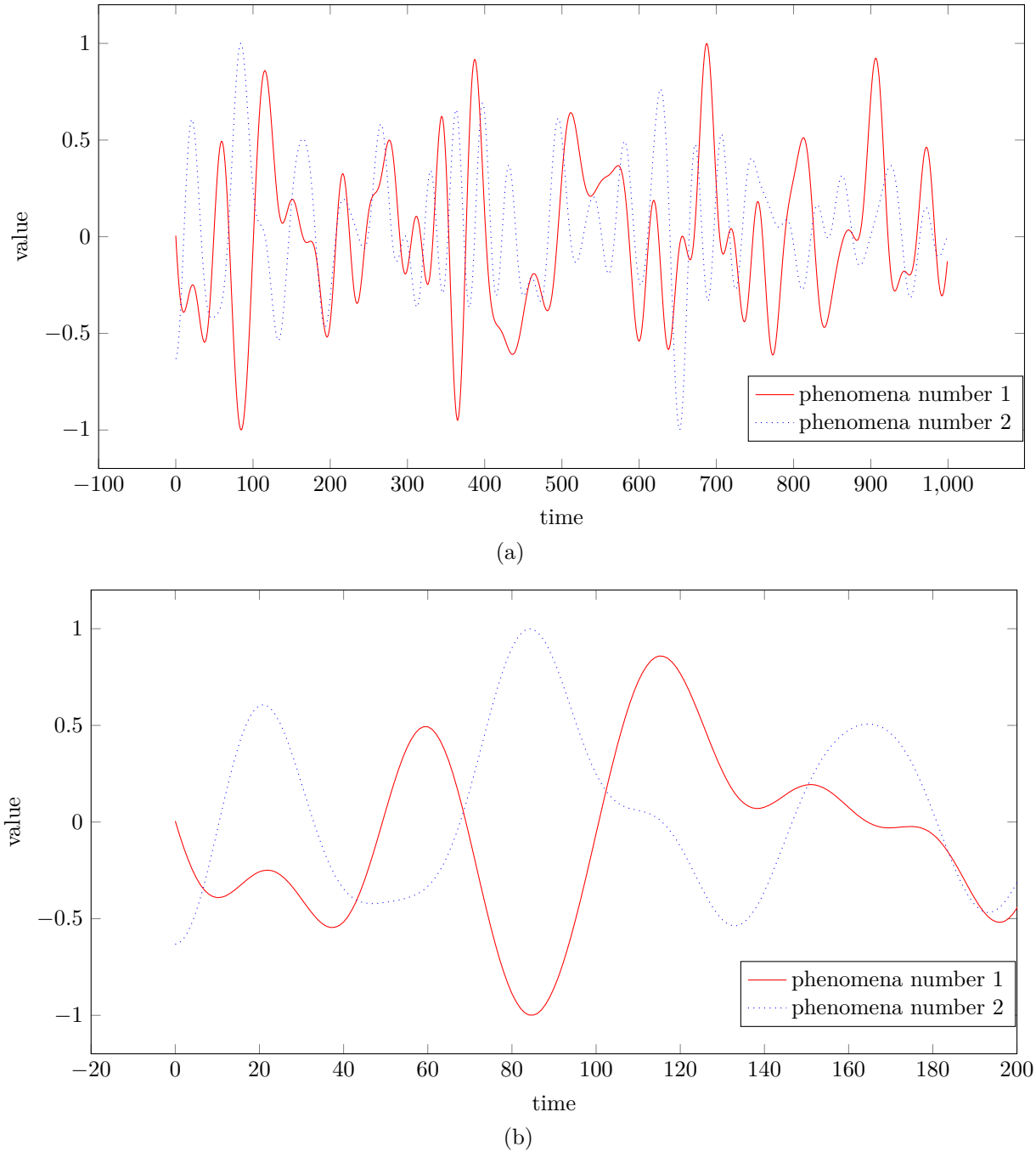


Figure 6.5: Phenomena: (a) in their entirety, (b) zoomed between $t = 0$ and 200.

Generation of Observation Histories

Each sensor follows one of the two phenomena, always the same one, and sends a noisy observation of the phenomenon, with Gaussian noise of standard deviation σ, i.e., $\mathcal{N}(0, \sigma)$. In this chapter, we study the impact of noise on a solution's ability to cluster sensors belonging to the same phenomenon. For this reason, we conduct a sampling of 500 simulations at constant intervals for $\sigma \in [0, 0.5]$ (keeping in mind that the values of the phenomena are bounded within $[-1, 1]$).

For a single simulation (i.e., a measurement noise level of σ), a new set of sensors' observations is generated. This set of observations is then used for all the compared methods.

The arrivals, departures, and transmission instants of the sensors are generated according to statistical laws. New sensors enter the environment over time, following a Poisson distribution with a parameter of $\lambda = 0.1$, and each of them follows one of the two phenomena with equal probability. The duration of a sensor's stay in the environment follows an exponential distribution with a parameter of $\mu = 0.01$. While in the environment, a sensor transmits observations following a Poisson distribution with a parameter of $\gamma = 1$.

We terminate the simulation at $t = 1000$. To avoid sensors that only have zero common definition intervals with other sensors (which can occur initially when few sensors are present), we consider only sensors that are still active after $t = 200$.

Given that the average duration of a sensor in the environment is $\frac{1}{\mu} = 100$, it is noteworthy that there are a considerable number of pairs of sensors with zero overlapping definition intervals.

From this scenario, we analyze the observation histories, where an observation history is defined as the set of observations generated by a sensor during its time in the environment. Our objective is to group sensors that track the same phenomenon.

4.2 Kriging Parameter Settings

The kriging requires fitting the experimental variogram to the variogram model. We have chosen the Gaussian model defined in Eq. (6.1). In the survey [PWG13], it was established that the choice of variogram model is relatively unimportant compared to the parameters associated with this model.

In our simulations, we used the Pykrige package in Python, which we utilized to create

kriging interpolations. This module can estimate the parameters of nugget n, sill s, and range r based on a given variogram model. However, since the observation histories are randomly generated with random measurement noise, the parameter estimation was not always accurate. In some cases, the parameter estimation led to very strong variations in the interpolation (e.g., small range r), while in other cases, it resulted in a nearly linear interpolation (e.g., very large range r).

To address this issue, for a given simulation, assuming that all compared observation histories have the same underlying form (since they are generated using the same random laws), they should be interpolated with the same variogram. To achieve this, for a given simulation, we fix the parameters n, s, and r that will be the same for all observation histories.

For one simulation, for each observation history i, we estimate the triplet of parameters n_i, s_i, and r_i using the fitting function provided by the PyKrig package. Consequently, for each parameter, we define the value for the variogram model across all observation histories in the simulation as the median value of the estimated parameters. For example, we set the range r as the median value obtained from the set of estimated ranges r_i, and this value is chosen for all observation histories.

4.3 Evaluation of the Similarity Metric

Let's begin by evaluating our choice of similarity metric based on interpolation. We compare our method to the state of the art and show its superiority.

4.3.1 Comparative Distance: Jaccard Distance

According to the state of the art presented in Chapter 3 Section 3, the only similarity metric that can compare the observation histories considered here is the one based on value sets. In particular, only the Jaccard distance can be directly applied to our case, and is proposed in [APM09].

The Jaccard score counts the proportion of common values between two value sets. If θ_1 and θ_2 represent two value sets of observations from an observation history, then the Jaccard distance is defined as:

$$d_{\text{Jaccard}}(\theta_1, \theta_2) = 1 - \frac{|\theta_1 \cap \theta_2|}{|\theta_1 \cup \theta_2|}$$

In our scenario, the values of observations are continuous. To allow for common values between two observation histories, we discretize these values. While evaluating the performance of the Jaccard metric (detailed in the performance evaluation) for time steps of 0.001, 0.005, 0.01, and 0.05, we select a time step of 0.005 as the visually optimal choice (with little difference from the time steps of 0.001 and 0.01).

4.3.2 Performance Evaluation

Methodology for Similarity Metric Comparison

We evaluate our distance based on interpolation and mean magnitude difference in comparison to the Jaccard distance. The objective of such distance metric is to give a large distance for sensors that follow different phenomena and a small distance for those that follow the same phenomenon.

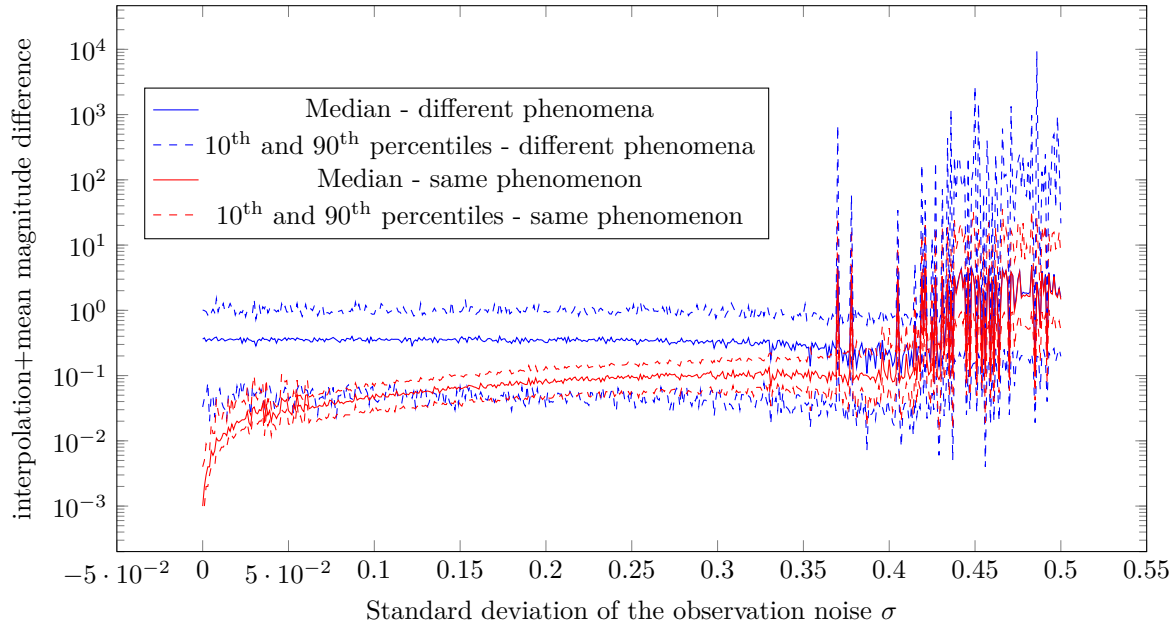
For each simulation and chosen distance metric, we retain all the existing distances computed between pairs of sensors. We categorize these distances into two groups: those calculated between sensors that follow the same phenomenon and those between sensors that follow different phenomena. It is worth noting that in the majority of cases (on average 80%), we cannot calculate a distance because the common definition interval durations are null.

For each distance metric, as shown in Fig. 6.6, we present the median distances along with the 10^{th} and 90^{th} percentiles, both for cases where distances are computed between sensors belonging to the same phenomena and when they are computed for sensors belonging to different phenomena.

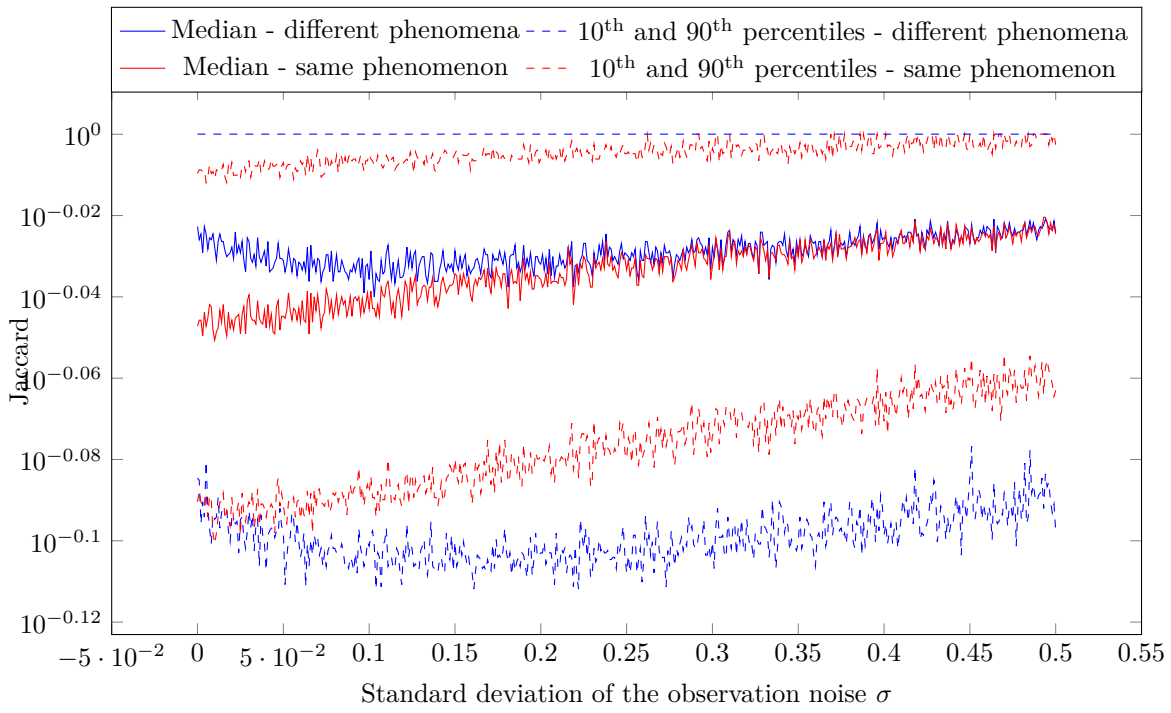
Our objective is to obtain a distance metric where sensors following the same phenomenon have relatively smaller distances compared to sensors tracking different phenomena. Thus, we aim for medians that are significantly apart and for intervals between the 10^{th} and 90^{th} percentiles to have minimal overlap.

Criticism of the Jaccard Distance

The Jaccard measure counts the number of times the observations are exactly the same within a range of 0.005 in a possible value space of $[-1, 1]$. Since the observations are taken at different time points, even if the observation histories follow the same phenomenon, there is little chance that they transmit exactly the same observation value. This observation is reflected in Fig. 6.6(b). We observe a difference of 5.2×10^{-2} between the two



(a)



(b)

Figure 6.6: Median, 10th and 90th percentiles of distances between sensors either from the same or different phenomena.
(a): Using our metric based on interpolation and mean amplitude difference.
(b): Using the Jaccard distance.

median distances for zero noise, which quickly drops to 8×10^{-3} for $\sigma = 0.2$. Similarly, the range of distances between the 10^{th} and 90^{th} percentiles overlaps significantly. In conclusion, the Jaccard measure fails to differentiate between pairs of sensors that follow the same phenomenon from those that follow different phenomena.

Discussion on the Performance of our Similarity Metric

On the other hand, using our interpolation and average magnitude difference distance yields significantly better performance, as shown in Fig. 6.6(a). For zero noise, the median distance is 1×10^{-3} when the sensors are following the same phenomena, whereas it is around 0.363 when the phenomena are different. Up to a noise level of $\sigma = 4 \times 10^{-2}$, the percentile intervals do not overlap, indicating that 90% of the distances for those following the same phenomenon are lower than 90% of the distances calculated between different phenomena.

Starting from $\sigma = 0.16$, the interval containing 80% of the median values (between the 10^{th} and 90^{th} percentiles) when following the same phenomenon is encompassed within the interval when following different phenomena. This implies that the distances between sensors belonging to the same phenomenon are mixed with the distances when the sensors do not belong to the same phenomenon. This phenomenon occurs due to the substantial dispersion of distances when sensors follow different phenomena. For instance, at $\sigma = 0.16$, the $10\% - 90\%$ interval size is 0.87 (compared to 0.06 when sensors track the same phenomenon).

The medians still remain relatively far apart. For example, at $\sigma = 0.38$, the median distance between sensors following the same phenomenon is 9.1×10^{-2}, compared to 0.249 when they follow different phenomena.

From $\sigma = 0.38$, we observe that the distances become much more variable, reaching maximal amplitudes of 10^3, which is unrealistic for observations within a constrained value range $[-1, 1]$ with a noise standard deviation lower than 0.5. This phenomenon is due to the fragility of the interpolation method, which struggles to interpolate accurately when points close in time are relatively far apart due to high noise levels.

4.4 Evaluation of the Clustering Solution

We now proceed to assess the performance of our clustering solution based on the constructed similarity metric. We develop a concurrent clustering method that aims to

minimize the maximum difference between interpolations within the same cluster. We demonstrate the superiority of our solution, which relies on the minimization of average differences.

4.4.1 Comparative Solution: Limiting the Maximum Distance Within a Cluster

Our solution is built to group sensors by seeking similar mean behaviors, and consists of two main components. Firstly, it employs a distance metric based on the interpolation of observation histories through kriging. This metric measures the difference in average magnitude between two interpolations over their common definition interval. Secondly, it utilizes an iterative hierarchical clustering method, merging clusters iteratively by selecting the smallest inter-cluster distance. This inter-cluster distance is defined as the weighted average of distances between sensors (using our distance definition) where the weights correspond to the duration of common presence between the compared sensors.

We propose to compare our solution to an approach extracted from the literature, specifically the solution proposed in [CKJ05; LWP07; TM06]. In these references, the sensors transmit observations at exactly the same time points. Two sensors are defined as similar if the maximum amplitude difference between their observations does not exceed a threshold. The sensors are grouped so that all pairs of sensors are similar within a group.

Itis essential to note that the assumptions underlying these references differ from those we have outlined in this chapter, which is why we propose a comparative solution inspired by these references. To facilitate a meaningful comparison between our approach and the one proposed in the literature, we retain some aspects of our methodology. The aim of this comparative solution (the same goal of the state-of-the-art proposal) is to restrict the maximum magnitude difference between all sensors interpolations within a cluster.

Hence, for the definition of distance, we maintain the same kriging-based interpolation method for observation histories while redefining the distance between two interpolations. Additionally, this new approach also relies on hierarchical clustering but employs a different linkage method.

Distance Based on Max Magnitude Difference

For this comparative solution, we retain the same interpolations made by kriging for this distance definition. The new distance is defined by the maximum deviation between

the two interpolations. According to the duration of the common definition interval $[a(i,j), b(i,j)]$ defined in Eq. (6.5), the distance $d_{\text{interp-max}}(i,j)$ is defined as follows:

$$d_{\text{interp-max}}(i,j) = \max[|\hat{\theta}_i(t) - \hat{\theta}_j(t)|, t \in [a(i,j), b(i,j)]]$$

Complete Linkage Method for Hierarchical Clustering

We choose to keep the same clustering algorithm as the solution proposed in this chapter and utilize the complete linkage [GR17]. This linkage method defines the distance between two clusters as the maximum existing distance between each pair of objects from different clusters:

$$D(I,J) = \max\{d(i,j), d(i,j) \neq None, i \in I, j \in J\}$$

If clusters I and J are merged, the distance between this new cluster and any other cluster K is defined as the maximum distance between I and K, and J and K:

$$D(I+J,K) = \max\{D(I,K), D(J,K)\}$$

In cases where one of these distances is $None$, the other one is used, and if both are $None$, the distance remains $None$.

Since, iteratively, the sensors group together while adhering to this rule, when clusters I and J are merged, this distance is greater than any distances between sensors belonging to the same cluster. For any cluster K present at this stage (including cluster $I+J$), it holds that $\forall k_1, k_2 \in K, D(I,J) \geq d(k_1,k_2)$.

Furthermore, the distance between sensors is determined by the maximum magnitude difference between their interpolations. Ultimately, the distance between two clusters that have just merged establishes a threshold. This threshold ensures that, for any pair of sensors within the same cluster, the amplitude difference between the interpolations does not exceed this distance.

Overall, following the initial principle outlined in the literature, the method we are developing limits the maximum magnitude difference between any pair of sensors belonging to the same cluster.

4.4.2 Threshold of 5 Clusters for the Agglomerative Hierarchical Clustering

We need to define a stopping condition for the hierarchical clustering algorithm to stop merging two clusters.

The solutions we are comparing are both based on the ascending hierarchical clustering method, and we want to compare them on the same evaluation ground. Since the distances, as well as the linkage methods, are different, we do not fix a distance-based threshold.

Therefore, we choose to opt for a threshold linked to the maximum number of clusters. The ideal threshold is to obtain 2 clusters, with, in the best case, one cluster containing the sensors following the first phenomenon, and the second containing those following the second phenomenon. However, as visible in the previous performance section, there are some pairs of sensors that follow different phenomena and have a very low distance: the values they return and that are interpolated are very close on their common definition interval. In this case, they would early be grouped. These groups formed by sensors following a different phenomenon would, a priori, have less similarity with sensors following one of the two phenomena. Therefore, we deliberately increase the desired number of clusters to isolate these small groups of sensors that were grouped together by mistake.

It is worth noting that in most cases, hierarchical clustering of sensors works well. These precautions are taken due to the large number of simulations and the random dimension of these simulations.

Thus, we arbitrarily impose the minimum number of clusters to be 5. In this case, this choice is not optimal (as explained, the optimal number is 2 clusters), but it is a compromise to obtain sufficiently consistent groups while allowing the isolation of groups of sensors that were initially grouped incorrectly.

It should be noted that in a more performant real-world implementation, one would prefer to apply a fixed or adaptive distance threshold.

4.4.3 Performance Evaluation

Methodology for Clustering Comparison

To assess the performance of a clustering solution, we evaluate the clustering results with respect to the true membership of sensors to the corresponding phenomena, called ground truth. For this purpose, we employ the Rand index proposed in [Ran71]. This metric quantifies whether, on average for a pair of objects, the two clustering solutions make the same choice. For any pairs of sensors, if both clustering solutions separate the two sensors or assign them to the same group, the score is 1; otherwise, the result is 0. The average is calculated over all pairs of objects to obtain a similarity score between 0 and 1.

Mathematically, let $(i)_{1\leq i\leq N}$ be the set of sensors, $(I)_{I\in\mathbb{I}}$ be a set of clusters (or a partition) created by one clustering method, and let the ground truth represent another clustering method $(J)_{J\in\mathbb{J}}$. Then, the similarity between $\mathbb{I}$ and $\mathbb{J}$, denoted by $c(\mathbb{I},\mathbb{J})$, is calculated as:

$$c(\mathbb{I},\mathbb{J}) = \frac{\sum_{i<j}^{N} \epsilon_{i,j}(\mathbb{I},\mathbb{J})}{\binom{N}{2}}$$

where

$$\epsilon_{i,j}(\mathbb{I},\mathbb{J}) \quad = \quad \begin{cases} 1 \text{ if there exist } I \in \mathbb{I} \text{ and } J \in \mathbb{J} \text{ such that } i \text{ et } j \text{ are both in } I \text{ and } J \\ 1 \text{ if there exist } I \in \mathbb{I} \text{ and } J \in \mathbb{J} \text{ such that } i \text{ is in both } I \text{ and } J \text{ while } j \\ \text{is in neither } I \text{ or } J \\ 0 \text{ otherwise} \end{cases}$$

We compare two clustering solutions, that are based on two paradigms: one that groups by seeking similar mean behaviors, and the other that groups by ensuring that within a group, the maximum deviations between the interpolations are limited. We used the same hierarchical clustering method for both solutions, setting the stopping rule to 5 groups to ensure fully comparable results. In Fig. 6.7, the clustering performance is visualized while varying the sensor measurement noise.

Comparison Results

Since both solutions rely on the same interpolations, when the noise becomes too significant ($\sigma \geq 0.38$), the performance are no longer reliable, and the similarity tends to converge towards random clustering performance, i.e. 0.5.

We propose to visualize the trends in the performance of each method using linear regression, which we perform on noise levels $\sigma < 0.3$. The performance interpolation of a clustering method concerning a standard deviation σ is denoted by $f_{\text{clustering method}}(\sigma)$, thereby yielding the following regression equations:

$$\begin{aligned} f_{\text{interp-mean-tuned}}(\sigma) &= -0.82\sigma + 1.00 \\ f_{\text{interp-max-complete}}(\sigma) &= -0.70\sigma + 0.88 \end{aligned} \tag{6.9}$$

The method based on the maximum and complete linkage relies on interpolation for distance calculation and seeks similarities in the definable interpolations. However, since the observations are not taken at the same time points, there is measurement noise, and

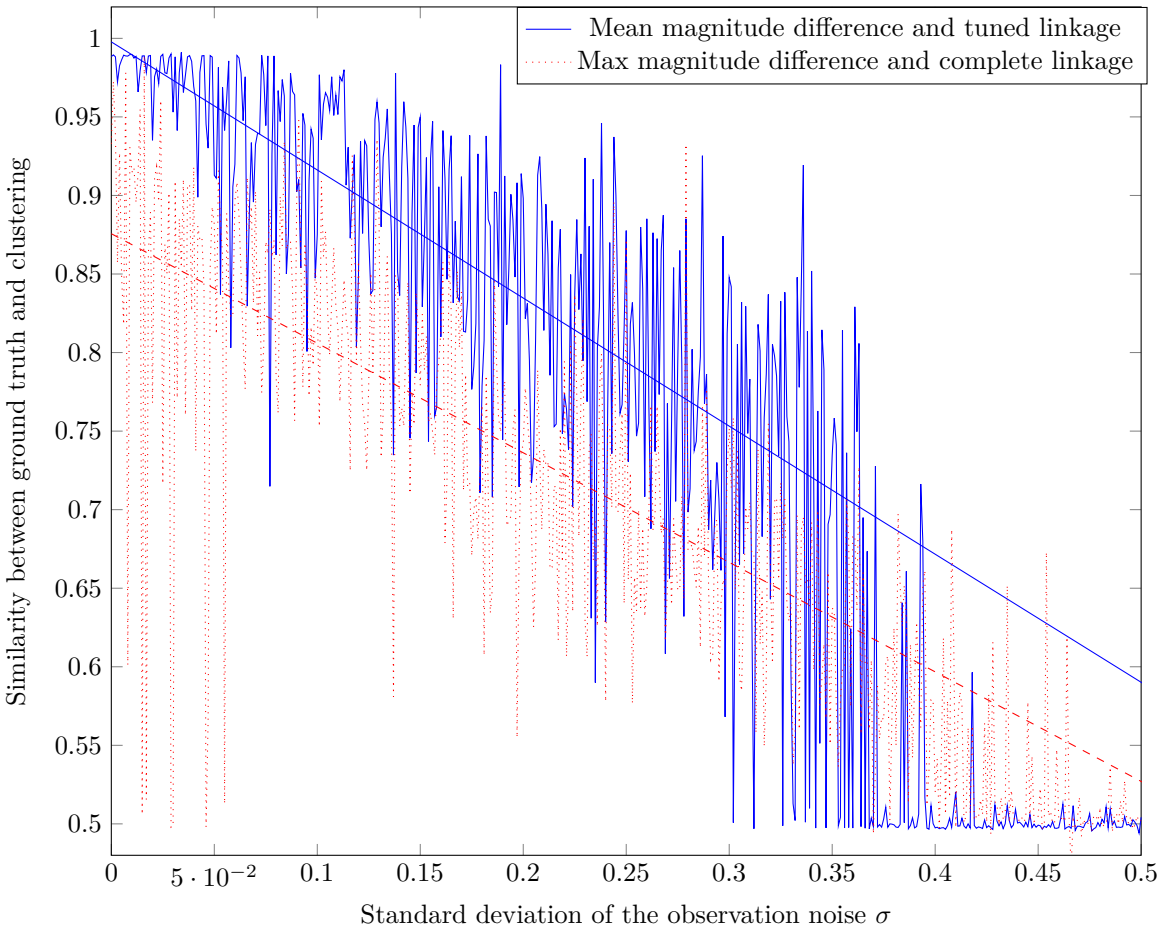


Figure 6.7: Similarity between ground truth and a clustering based on the observations histories. Linear regression of performance for $\sigma < 0.3$.

the two phenomena are very close and variable, making the choice of maximum difference between all pairs of interpolations not an effective strategy. Based on the regressions proposed in Eq. (6.9), we can assert that there is at least a superiority of 9.7×10^{-2} in similarity for $\sigma < 0.3$ in favor of the method based on the mean, representing a 17% minimum improvement.

In our previous conclusions in Section 4.3.2 regarding the evaluation of the distance between sensors by calculating the mean difference of interpolations, we concluded that from $\sigma = 0.16$ there was a significant overlap in distances when sensors followed the same phenomenon and when they followed a different one. However, we can see that the clustering method remains relatively robust to this, offering (based on the regression results) a similarity of 0.87 between the clustering result and the ground truth when $\sigma = 0.16$.

However, there is relatively high variability in the performance of the clustering methods. We interpret this variability as being influenced by the strict choice of 5 clusters and the use of hierarchical clustering. Indeed, dividing the set of observation histories into 5 clusters by default, compared to the reference of 2 clusters, influences negatively the results.

5 Conclusion and Perspectives

In this chapter, we proposed a method for grouping sensors based on their observations. Our approach was designed to handle a pessimistic scenario where sensors come and go over time, sending irregular and noisy observations of possibly different phenomena.

Firstly, we developed a similarity metric that accommodates irregular and non-synchronous observations. We demonstrated that our distance metric effectively characterizes pairs of sensors when they follow the same phenomenon versus a different one, outperforming the Jaccard distance.

Additionally, we use a hierarchical clustering method that relies on our similarity metric, accounting for variable and possibly null coexisting interval between two sensors. Our proposed approach aims to find groups of sensors that are close on average. Compared to a method inspired by the state-of-the-art, which seeks to limit the maximum difference between observations within a cluster, our solution shows at least a 17% improvement in terms of clustering performance.

However, there are still further analyses to be conducted, as we have explored only

a portion of potential solutions. For example, alternative interpolation methods or different comparison metrics between two interpolations could be considered. Furthermore, while we implemented hierarchical clustering methods, there is a vast array of clustering techniques available that could be explored to enhance our approach.

Still, our findings in this pessimistic scenario highlight the feasibility of identifying groups of similar sensors by analyzing their observations. Such an analysis enables the application of observation collection schemes based on this similarity. For instance, the Asynchronous 2-Level Round-Robin method presented in Chapter 5 could be utilized for each identified cluster. An experimental study combining the three components of an observation collection scheme would further validate this approach of observation collection scheme.

CONCLUSION

Chapter 7

Conclusion and Research Directions

1 Summary of our Works

Currently, there exists a plethora of mechanisms aimed at reducing sensor power consumption. In this thesis, we have introduced a novel energy-saving mechanism based on the efficient collection of sensor observations through similarity. This approach relies on deploying a large number of sensors, a common paradigm experiencing a resurgence due to the development of IoT-enabled networks, referred to as MIoT.

We have devised the concept of an observation collection scheme, which relies on the principle of similarity to reduce the volume of observations transmitted by sensors and thereby preserve their energy resources. To expound on this method, we conducted the first thorough investigation of the existing literature, representing a proposal into three primary components. Specifically, the *similarity metric* enables the assessment of similarity between sensors based on available information. Using this similarity metric, the *covering subset algorithm* generates one or more subsets of sensors, each capable of fulfilling monitoring requirements. Finally, the *activation allocation method* distributes the observation load among sensors. For each of these components, we conducted a thorough examination of existing proposals, significantly influencing the development of our contributions.

• For the *similarity metric* component, we have concluded that it is crucial to rely on sensor observations-based similarities, this allowing for the consideration of realistic and complex environments. However, we found that there were no suitable proposals in the current literature to evaluate similarity between sensors based on their observations when transmissions are not synchronized.

In response to this, we proposed in Chapter 6 a similarity metric that utilizes kriging as an interpolation tool for the observations and evaluates the mean magnitude differ-

ence between interpolations over the common time intervals between two sensors. We demonstrated the superiority of our approach over the state-of-the-art Jaccard distance. However, we also observed that under highly uncertain conditions (significant noise), the evaluation of similarity did not yield satisfactory results, as the interpolation method was no more accurate.

• Regarding the *covering subset algorithm* component, our state-of-the-art study led to the conclusion that it is more appropriate to develop algorithms that involve the entire set of sensors as multiple covering subsets rather than seeking a single covering subset as the solution. Additionally, we found that a clustering-based approach is more suitable than the disjoint covering approach. The clustering method can handle the heterogeneity of similarity links and create sensor groups of varying sizes, making it more suitable in our context of MIoT deployment.

In Chapter 6, we proposed a hierarchical clustering method and adapted the linkage method to account for the varying common presence time between two sensors. We compared our solution to the most common state-of-the-art method that aims to form groups by minimizing the maximum observation deviations within each group. The simulation results demonstrated that attempting to limit the maximum deviations between sensors within the same group led to decreased performance in our context characterized by high variability, noisy observations, and the possibility of sensors following different yet closely related phenomena. In contrast, our clustering-based solution, which focuses on grouping sensors with similar average observations, proved to be more reliable under such challenging scenarios.

• Finally, concerning the *activation allocation method*, the most suitable existing solutions from the literature propose a round-robin-based approaches, where covering sensor subsets activate at constant time slots, taking turns to transmit instead of all transmitting simultaneously. However, these literature methods tend to be quite rudimentary and not well-suited to the uncertainties encountered in a MIoT deployment.

We then proposed two solutions, each discussed in separate chapters. The goal of these approaches is to build upon a clustering solution. Once a cluster of similar sensors is identified, we aim to distribute the observations load evenly among the sensors. Specifically, we sought to receive the same quantity of messages regardless of the number of similar sensors within the cluster. Since the number of sensors and evaluated similarity can vary over time, the groups may also change over time. As a result, we focused on adapting the activation periods relative to these changes, minimizing the costs associated

with period adjustments.

In a preliminary version (Chapter 4), we presented a formalism for a function to update the activation periods. Our proposed method allows for the dynamic inclusion of new sensors and ensures strictly regular reception, controlled by the time interval between receptions and the number of sensors transmitting in round-robin mode. This study highlighted the trade-off between tracking quantity and system lifetime, which is controlled by the two parameters of our function. We demonstrated that to achieve the best balance between these objectives, it is generally preferable to activate only a subset of sensors in the round-robin process. Activating all sensors in round-robin mode becomes costly when dealing with a large set of them, as including/excluding a sensor requires adapting the activation periods of all other sensors.

In Chapter 5, we propose a method that ensures a constant quantity of observations over time, while efficiently including and excluding sensors with minimal period change costs. We also relax the strict periodic reception constraint imposed by the previous method since achieving strict temporal synchronization for a fleet of sensors is costly in practice. Through our solution, we demonstrate that minor changes to individual sensor periods and the relaxation of strict message reception have minimal impact on tracking accuracy. Additionally, our approach requires significantly fewer period adjustments, which is advantageous in terms of sensor power consumption and network downlink overhead. We present theoretical results on the estimation of tracking quality based on the function parameter, helping to optimize this parameter effectively.

This dissertation has contributed to formalizing a novel energy-efficient mechanism for the IoT, enabling more efficient transmission of sensor observations. This method is applicable within the context of MIoT, where highly constrained devices are widely distributed, for instance in everyday objects. While this mechanism complements existing energy-efficient approaches for IoT, it extends the capability to encompass a broader range of sensors, including those with extremely limited capacities transmitting on highly constrained networks — a feat not achievable by all existing energy-efficiency mechanisms.

2 Limitations and Perspectives

2.1 Need to Address All the Fixed MIoT's Points of Interest

We have developed proposals that address most of the points of interest for a MIoT deployment that we initially defined in Chapter 2 Section 5. However, our solutions still fall short of satisfying all the specified expectations.

Regarding the activation allocation methods presented in Chapter 4 and Chapter 5, we extensively worked on handling variations in the number of sensors, with the aim of minimizing downlink transmissions. In Chapter 5, we tackled the challenge of downlink packet loss, wherein the solution no longer requires two consecutively well-received orders from the terminal. For the similarity metric and the covering subset algorithm, our work in Chapter 6 also attempted to account for significant variations of the number of present sensors. By utilizing a data-based similarity approach, we effectively managed complex and realistic environments. Our solution handled clock drift issues and noisy sensor measurements. The clustering structure allowed for easy isolation of poorly placed sensors, although this aspect was not thoroughly explored in the simulations.

However, notably for the similarity metric and the covering subset algorithm components, we did not account for potential variations in similarity between sensors over time. Such situations may arise when a sensor becomes corrupted, and we need to exclude it from its group of similar sensors, or when the environment undergoes changes, prompting the grouping or separation of similar sensor groups. One possible approach to address this limitation is to employ online change detection methods, such as using sliding windows to identify evolving similarities over time.

Regarding the activation allocation methods, we made the assumption of perfect message reception, but in reality, observations are not always received, and there can be a need for including this in the solution. We will discuss this aspect further in Section 2.3.3 but adapting the quantity of sent messages to the proportion of successfully transmitted messages can be a solution to maintain a desired quantity of received observations.

2.2 Benefits of Establishing a Testbed

The work presented here remains at the simulation stage, representing a proof of concept. Similar methods from the state of the art have already demonstrated the relevance of such approaches in testbeds, showing that using the similarity principle can significantly reduce

the number of transmitted messages without significantly affecting tracking quality.

To further validate our approach, it would be pertinent to conduct a real-world study of this scenario. While we attempted to create a realistic scenario for phenomena, sensor transmissions, and observation quality, our similarity and clustering solutions have, so far, been evaluated on synthetic data, which does not completely validate the proposal. A comparative study of multiple similarity and clustering solutions on real data would provide concrete conclusions on the optimal choices for each component. This real-world testbed would allow us to observe the performance of our proposed solutions in practical scenarios, thus strengthening the validity and applicability of the methodology.

Up to this point, the experiments conducted in the literature have primarily focused on recognizing the potential for reducing sensor observations without implementing a concrete policy for changing sensor activation periods. Applying such policy in an experiment would inevitably lead to technical challenges. For instance, issues related to real traffic congestion in both uplink and downlink could arise. Additionally, it is crucial to assess the real-time reprogrammability of sensor activation periods after their activation, especially in the context of LoRaWAN Class A devices.

2.3 Generalizing the Problem: Managing Heterogeneous Sensors

Throughout this thesis, we assumed a single network with standard sensors transmitting standardized observations. However, in reality, we can encounter different development phases, with energy-linked sensors transmitting over wired networks and coexisting with other deployments using different sensor brands and networks.

2.3.1 Different Networks, Different Data Formats

Today, multiple types of networks coexist without the possibility of interconnection. Moreover, they transmit data in formats that are different from one sensor to another. These challenges can be characterized as interoperability issues and are theoretically addressed in [TDT07; Tol03] as the first two levels of system interoperability. These challenges are not fully addressed yet and are critical for the realistic implementation of the solutions we consider.

2.3.2 Integration of Sensors Connected to an Energy Source

It is crucial to distinguish between sensors that are linked to an energy source and those that are not. For sensors connected to an energy source, the number of transmissions does not impact their lifespan. Thus, they can transmit observations more frequently without energy constraints.

Integrating these energy-source-connected sensors into a dense sensor fleet requires the application of similarity metric components and search for covering sets to establish links of similarity with other sensors. These similar sensors (powered by batteries) can preserve their energy more effectively and serve as a means of validating observations transmitted by sensors with an infinite energy source. They can take over if the sensor connected to an energy source becomes corrupted.

2.3.3 Diverse Transmission Efficiencies

The transmission performance of sensors (packet delivery ratio for instance) varies based on the transmission mode and sensor location.

Integrating such parameters into observation collection management policies is possible. At least two types of strategies can be envisaged: those based on pure energy efficiency and those based on pure load distribution.

One approach involves optimizing the overall consumption by prioritizing sensors with cost-effective transmissions. On the other hand, this strategy may overconsume these efficient sensors, resulting in the depletion of batteries, leaving only the lower-quality sensors operational.

Another choice is to distribute the transmission load among all sensors at all costs. This involves increasing the message sending rate of sensors with lower transmission probabilities to ensure an adequate message quantity is received, albeit potentially accelerating their energy depletion. While this scenario could lead to a shorter system lifespan due to increased consumption, it does not excessively burden the more reliable sensors.

2.3.4 Varying Sensor Accuracy

The brand and quality of a sensor necessarily influence the precision of the measurement tool. Even with standardized objects, precision can vary from one object to another. It is relevant to handle this variability in sensor management policies.

In a broad sense, under assumptions of Gaussian noise (as considered in this thesis),

increasing the number of observations enhances precision. Consequently, for a given precision requirement, a more accurate sensor would need to transmit fewer observations than a less accurate one.

Similarly, akin to the diversity in transmission efficiency, we can outline two primary strategies: those emphasizing energy efficiency and those centered on load distribution.

One strategy could prioritize sensors with high accuracy, as fewer transmissions would be necessary to achieve a certain level of tracking precision. However, once these high-precision sensors deplete their batteries, only the remaining less-precise sensors would continue functioning, inevitably leading to a performance decline.

Conversely, we can pursue a strategy of precision balance, wherein less accurate sensors transmit more frequently to attain the same accuracy level as the more accurate sensors. This approach would allow precise sensors to assist in calibrating less accurate ones, which are more susceptible to drift and fixed errors. Nonetheless, this solution is less energy-efficient due to the increased transmission demands on the sensors.

APPENDIX

Appendix A

Résumé en Francais

Les objets connectés, ou "Internet of Things" (IoT), sont couramment utilisés pour surveiller diverses grandeurs physiques. Dans l'approche novatrice du Massive IoT (MIoT), on imagine un déploiement massif de capteurs à faible coût alimentés par batterie intégrés à des objets du quotidien, par exemple sous chaque plaque de faux plafond dans un bâtiment, dans le but de réduire les coûts de déploiement et de maintenance. Cette thèse se concentre sur le développement de mécanismes visant à réduire la consommation d'énergie des capteurs, dans le but de prolonger la durée de vie de ce type de déploiement IoT. Plus précisément, nous étudions dans cette thèse les méthodes qui s'appuient sur la similarité entre capteurs pour gérer la collecte d'observations de capteurs de manière plus efficace.

Dans Chapter 2, le scénario de déploiement Massive IoT est défini, mettant en lumière les hypothèses relatives aux capteurs, au réseau, et à l'environnement étudié. Les capteurs sont considérés comme étant déployés en grand nombre, ce qui implique qu'ils sont proches les uns des autres et susceptibles de retourner des observations très similaires. L'objectif est d'identifier ces liens de *similarité* entre les capteurs pour gérer efficacement leurs transmissions. Par exemple, une approche consiste à recevoir uniquement les messages d'une sous-partie de l'ensemble des capteurs, que l'on appelle un ensemble *couvrant*. Les autres capteurs restent alors en mode veille, tout en garantissant que les exigences de précision de suivi de l'environnement sont respectées. Pour répondre aux contraintes des capteurs et du réseau, une solution de gestion des observations des capteurs est définie par la mise à jour de la période de transmission d'un capteur à la suite de sa transmission de message.

Ensuite, en Chapter 3, une étude est menée sur les méthodes existantes issues de l'état de l'art. Ces méthodes partagent l'application du principe de similarité entre les capteurs pour réduire la quantité de messages transmis par la flotte de capteurs. Cette étude regroupe des travaux provenant de différents domaines de recherche. À ce jour, aucune étude de cette envergure n'a été entreprise dans la littérature. L'un des domaines

étroitement liés vise à réduire le volume de messages transmis en utilisant des stratégies de planification des observations des capteurs. Ces travaux établissent des liens de similarité entre deux capteurs qui transmettent des observations suffisamment similaires. À partir de ces liens, des sous-groupes de capteurs couvrants sont créés, de sorte que chaque sous-groupe englobe tous les capteurs en utilisant l'extension de leurs liens de similarité. Au lieu de permettre à tous les capteurs de transmettre en permanence, les sous-groupes couvrants transmettent en alternance. Un autre domaine étroitement lié à notre problématique vise à obtenir une couverture spatiale cible en utilisant un minimum de capteurs pour répondre à cette exigence de surveillance. Par exemple, les capteurs couvrent une zone grâce un rayon de couverture, et on cherche à couvrir un pourcentage minimum de l'environnement en utilisant un minimum de capteurs. Pour unifier l'ensemble de ces contributions, nous décomposons ce type de méthode en trois composantes principales : la métrique de similarité, l'algorithme de sous-ensemble couvrant, et la méthode d'allocation d'activations. La **métrique de similarité** est une métrique réelle qui quantifie la proximité entre les capteurs, en se basant sur les informations connues des capteurs. À partir de cette métrique de similarité, l'**algorithme de sous-ensemble couvrant** construit un ou plusieurs sous-ensembles de capteurs, où chaque sous-ensemble garantit la satisfaction des exigences de surveillance de la grandeur physique. Enfin, en s'appuyant sur les sous-ensembles de couverture, la **méthode d'allocation d'activations** définit comment répartir la charge des transmissions d'observations entre les capteurs. Pour chaque composant, nous présentons ce qui a été proposé dans la littérature, en identifiant les limites existantes.

Les chapitres suivants constituent les contributions de la thèse. En particulier, nous avons étudié la méthode d'allocation d'activation dans Chapter 4 et Chapter 5 ; la mesure de similarité et l'algorithme de sous-ensemble couvrant sont étudiés dans le chapitre Chapter 6. Nous présentons ainsi les conclusions et limites tirées de notre étude de la littérature, ainsi que nos propositions de réponse.

• En ce qui concerne la métrique de similarité, dans la littérature, elle est calculée soit à partir de la position géographique des capteurs, soit à partir des observations retournées par les capteurs. Notre étude conclut qu'il est préférable de s'appuyer sur une métrique basée sur les observations retournées par les capteurs, car elle permet de mieux décrire des environnements complexes. Les solutions existantes traitent un historique d'observations comme un ensemble de valeurs, ignorant la dimension temporelle des observations, ou supposent que tous les historiques d'observations sont effectués simultanément. Ces deux

approches présentent des limitations significatives, notamment la perte de précision en supprimant la dimension temporelle des observations, ainsi que la complexité en pratique de la synchronisation des capteurs pour obtenir des historiques d'observation simultanés.

Nous avons ainsi proposé en chapitre Chapter 6 une mesure de similarité basée sur les observations qui contourne ces limitations. À partir d'un historique d'observations quelconque, nous définissons une méthode interpolation de krigeage, prenant en compte d'éventuelles erreurs de mesure. La métrique de similarité entre deux capteurs est ensuite calculée comme la différence moyenne entre les interpolations sur leur intervalle de définition commun. Nous avons comparé notre solution à la métrique de Jaccard, qui évalue la proportion de valeurs similaires entre deux ensembles de valeurs (en ignorant la dimension temporelle), et nous montrons sa supériorité à minimiser la distance entre les capteurs suivant un phénomène similaire et maximiser la distance entre les capteurs suivant un phénomène différent.

• En ce qui concerne l'algorithme de sous-ensemble couvrant, les méthodes existantes proposent soit de rechercher un sous-ensemble couvrant, soit de partitionner l'ensemble des capteurs en sous-ensembles couvrants, soit de créer des clusters de capteurs similaires, où chaque capteur couvre les autres capteurs appartenant au même cluster. Nous avons conclu que pour s'adapter aux variations du nombre de capteurs et des similarités, il était préférable d'opter pour une structure de clustering.

Dans le chapitre Chapter 6, nous présentons une méthode de clustering hiérarchique en définissant la distance inter-cluster comme la distance moyenne entre les capteurs, pondérée par la durée de leur intervalle de définition commun. Cette approche accorde davantage d'importance aux paires de capteurs présents ensemble sur des durées plus longues, ce qui renforce la fiabilité de la distance calculée. Nous avons comparé cette solution (en utilisant notre métrique de similarité basée sur l'interpolation) a une méthode adaptée d'une référence de la littérature. Cette méthode de comparaison vise à faire des groupes de capteurs similaires, en limitant la différence maximale entre les interpolations de toutes paires de capteurs appartenant au même cluster. Nous montrons la supériorité de notre proposition en termes de capacité à grouper correctement des capteurs qui suivent un même phénomène. Ainsi, nous soulignons qu'il est plus pertinent d'évaluer les comportements moyennement similaires dans un environnement où les observations sont effectuées à des intervalles irréguliers et sont soumises au bruit.

• Enfin, en ce qui concerne la méthode d'allocation d'activations, les solutions de la littérature proposent soit d'utiliser un seul ensemble couvrant pour répondre aux besoins de

monitoring, soit de faire tourner sous forme de round robin strict les ensembles couvrants construits. Nous avons interprété comment ces méthodes sont appliquées dans notre réseau à fortes contraintes, c'est-à-dire sous la forme de la mise à jour de la période de transmission d'un capteur qui vient de transmettre. Nous avons conclu qu'il est préférable d'utiliser la totalité des capteurs disponibles dans la solution (méthodes round-robin), car cela nécessite moins d'ordres à donner aux capteurs et permet d'identifier plus facilement qu'un capteur devient aberrant en le comparant aux capteurs similaires.

Nous proposons deux méthodes qui permettent, au sein d'un groupe de capteurs similaires, de recevoir une quantité constante de messages par unité de temps, tout en prenant en compte que le nombre de capteurs varie au cours du temps. La première méthode, développé en Chapter 4, permet de recevoir des messages à intervalle constant cible, en utilisant un round-robin strict entre un nombre maximum de capteurs. Nous concluons qu'un compromis doit être trouvé entre la durée de vie du système et la précision du suivi, et nous montrons qu'il est préférable d'utiliser uniquement un sous-ensemble de capteurs en round-robin pour obtenir une solution optimale. Dans une seconde méthode, décrite en Chapter 5 nous assouplissons la réception stricte des messages pour la rendre plus facilement applicable, et nous proposons des mécanismes nécessitant un nombre limité d'ordres de changement de périodes pour maintenir la propriété de réception d'observations à intervalles de temps constants. La première solution présente des propriétés mathématiques d'optimalité en termes de précision. La seconde solution obtient presque les mêmes performances en expérimentation, tout en demandant un nombre bien moindre de changements de périodes.

Appendix B

Proofs of Bounds and Effectiveness in Chapter 4

In this part of the appendix, we provide the proofs for the propositions presented in Chapter 4. Firstly, we present the proof of the upper bound of the sample span for an effective allocation function in Appendix B.1. Next, in Appendix B.2, we demonstrate that the developed function $f_{M,\tau}$ is effective on the instants of period τ. Lastly, we establish bounds for the sample span of $f_{M,\tau}$ in Appendix B.3.

B.1 Upper Bound of an Effective Period Allocation function

For the analysis, we explicit the scenario Π, considering that the sensors entering the environment are indexed in ascending order of arrival. For the n incoming sensors in the environment, they are indexed as $(i)_{0\leq i\leq n-1}$, $t_0 < t_1 < \ldots < t_{n-1}$.

B.1.1 Preliminaries

Our objective is to characterize the sample span in terms of the number of period changes. To achieve this, we introduce the terms of number of period changes made by a sensor i during the monitoring duration using an effective period allocation function f, that we denoted by $r_i(f)$.

Proposition 8. *The sample span of a period allocation function f, effective over the instants of period τ is:*

$$L(f,\Pi) = \sum_{i=0}^{n-1} \left\lfloor \frac{e_i - c_e - c_r r_i(f)}{c_e} \right\rfloor \tag{B.1}$$

Proof. For a sensor i, it consumes energy during the initial activation (c_e) and also the period changes ($c_r r_i(f)$). Its remaining energy ($e_i - c_e - c_r r_i(f)$) is consumed over the instants of period τ. Hence, the number of activation on these time steps corresponds to $\left\lfloor \frac{e_i - c_e - c_r r_i(f)}{c_e} \right\rfloor$.

Since the activations on instants of period τ between sensors are disjoint and not discontinuous, the number of observations is exactly the sum of the observations made by each sensor at the instants.

□

Lemma 1. *Let f and g be 2 effective period allocation functions. If $\forall i, r_i(f) \geq r_i(g)$, then the sample span of g is greater than that of f : $L(f, \Pi) \leq L(g, \Pi)$.*

Proof.

$$r_i(f) \geq r_i(g) \implies \left\lfloor \frac{e_i - c_e - c_r r_i(f)}{c_e} \right\rfloor \leq \left\lfloor \frac{e_i - c_e - c_r r_i(g)}{c_e} \right\rfloor$$

This being true for all i, by summation and from Proposition 8, we have that $L(f, \Pi) \leq L(g, \Pi)$ □

B.1.2 Demonstration of Eq. (4.2) (Page 65)

Proof. We want to prove that all period allocation functions are upbounded by $\frac{\sum_{i=0}^{n-1} e_i - nc_e - (2n-1)c_r}{c_e}$.

We can represent an inequality in order to describe the sample span of a period allocation function:

$$\frac{(\sum_{i=0}^{n-1} e_i) - nc_e - (2n-1)c_r}{c_e} \geq \sum_{i=0}^{n-1} \left\lfloor \frac{e_i - c_e - c_r(1 + \mathbb{1}_{i>0})}{c_e} \right\rfloor$$

Hence, we need to prove that the minimum number of period changes is $1 + \mathbb{1}_{i>0}$.

In the proposition, it is assumed that the sensors do not come alive at the instants of period τ i.e.

$$t_i \not\equiv t_0[\tau] \tag{B.2}$$

Then, let us consider f an effective period allocation function. We will show that $\forall i, r_i(f) \geq 1 + \mathbb{1}_{i>0}$.

-$i = 0$. Necessarily, f change at least one time the period of the sensor 0. Then $r_0(f) \geq 1$.

-$i > 0$. Let us suppose $r_i(f) < 2$, then $r_i(f) = 1$. We note p its period of activation. Since f is effective, its first activation after initialization is on the time steps τ i.e. $t_i + p \equiv t_0[\tau]$.

But from (B.2), t_i is not on the instants of period τ, hence the period p is not a multiple of τ: $t_i + p \equiv t_0[\tau] \& t_i \not\equiv t_0[\tau] \implies p \not\equiv 0[\tau]$. Looking at the following activation $t_i + 2p$, we come across an absurdity. Thus $r_i(f) \geq 2$.

Hence, $r_i = 1 + \mathbb{1}_{i>0}$ is the minimum number of changes for each sensor. From Lemma 1, f has a lower sample span than a period allocation function with period changes equal to $r_i = 1 + \mathbb{1}_{i>0}$:

$$\begin{aligned} L(f, \Pi) &\leq \sum_{i=0}^{n-1} \left\lfloor \frac{e_i - c_e - c_r(1+\mathbb{1}_{i>0})}{c_e} \right\rfloor \\ &\leq \frac{(\sum_{i=0}^{n-1} e_i) - nc_e - (2n-1)c_r}{c_e} \end{aligned}$$

□

B.2 Effectiveness of $f_{M,\tau}$ Over the Instants of Period τ

We first introduce some tools to track the behavior of sensors over time, and then propose a proof of the effectiveness of the function $f_{M,\tau}$.

B.2.1 Sensor Representation Set

Definition 6. *A sensor i alive at time t is represented by the state $E_i(t)$, such that:*

$$E_i(t) := \begin{cases} e_i(t) \geq c_e : \textit{remaining energy at time } t \\ p_i(t) > 0 : \textit{activation period} \\ \delta_i(t) > 0 : \textit{time before sending a message after } t \end{cases} \tag{B.3}$$

$E_i(t) := \emptyset$ if the sensor i is not alive at time t (didn't initialized yet or is dead).

We denote by $\Pi(t)$ the set of alive sensors at time t, $|\Pi(t)|$ is the quantity of alive sensors at time t. For instance, we have that $i \in \Pi(t) \Leftrightarrow E_i(t) \neq \emptyset$.

We define $E(t)$, so that:

$$E(t) := (E_i(t))_{i \in \Pi(t)}$$

The formalization of the state $E(.)$ will help to prove that the function $f_{M,\tau}$ is effective, i.e. the activations of the sensors (except initializations) are on the instants of period τ.

First of all, we can define $E(.)$ until the first activation of the first sensor:

$$\begin{gathered} t < t_0, E(t) = () \\ E(t_0) = (E_0(t_0)), \& \; E_0(t_0) = \begin{cases} e_0(t_0) = e_0 - c_e - c_r \\ p_0(t_0) = f(\Pi(t_0)) \\ \delta_0(t_0) = f(\Pi(t_0)) \end{cases} \end{gathered} \tag{B.4}$$

The initial activation period of the sensor is defined by f. The sensor consumes energy for its first activation and the setting of its period: $c_e + c_r$. Then, from time t_0, the next activation occurs after a duration of $f(\Pi(t_0))$.

Moreover, $E(.)$ evolves over time, for each activation. We characterize the variations from state $E(t)$ to state $E(t + \Delta t)$.

To clearly track the updates of activation periods, we will assume that sensors change their activation period at most once between t and $t + \Delta t$. More generally, we consider Δt to be smaller than all activation periods:

$$\begin{aligned} \forall i, \quad & \Delta t \leq p_i \\ & \Delta t \leq f(.) \end{aligned} \tag{B.5}$$

We now characterize the evolution from t to $t + \Delta t$ under condition (B.5).

• If i is a sensor that is alive at time t, i.e. $i \in \Pi(t)$. Moreover, if the sensor does not activate between t and $t + \Delta t$, $\delta_i(t) > \Delta t$. Then, the periods and energies states don't change. The duration before the next activation at time $t + \Delta t$ is decreased by Δt:

$$E_i(t + \Delta t) = \begin{cases} e_i(t + \Delta t) = e_i(t) \\ p_i(t + \Delta t) = p_i(t) \\ \delta_i(t + \Delta t) = \delta_i(t) - \Delta t \end{cases} \tag{B.6}$$

• If i was already alive and activates between t and $t + \Delta t$: $0 < \delta_i(t) \leq \Delta t$. Then, it consumes an energy c_e. Moreover, the function f is used to determine the new activation period of the sensor. It will consume an additional energy c_r if the defined period is different from the current one. The sensor is represented in $E(t + \Delta t)$ only if it is alive at $t + \Delta t$ i.e. if it has enough energy to transmit again. In order to define a simple form of $E_i(t + \Delta t)$, we define the energy remaining in the sensor i at time $t + \Delta t$. $e_i(t + \Delta t) = e_i(t) - c_e - c_r \mathbb{1}_{f(\Pi(t+\delta_t)) \neq p_i(t)}$ ($\mathbb{1}$ is the indicator function).

Then:

$$E_i(t+\Delta t) = \left[\begin{array}{ll} \emptyset & \text{, if } e_i(t+\Delta t) < c_e \\ \left\{ \begin{array}{l} e_i(t+\Delta t) \\ p_i(t+\Delta t) = f(\Pi(t+\delta_t)) \\ \delta_i(t+\Delta t) = \delta_i(t) - \Delta t + f(\Pi(t+\delta_t)) \end{array} \right. & \text{, else} \end{array} \right. \tag{B.7}$$

• Finally, if a new sensor i comes alive between t and $t+\Delta t$, $t_i \in]t, t+\Delta t]$, $i \notin \Pi(t)$, $i \in \Pi(t+\Delta t)$. In this case, f defines the activation period of i and:

$$E_i(t+\Delta t) = \left\{ \begin{array}{l} e_i(t+\Delta t) = e_i - c_e - c_r \\ p_i(t+\Delta t) = f(\Pi(t_i)) \\ \delta_i(t+\Delta t) = t_i - (t+\Delta t) + f(\Pi(t_i)) \end{array} \right. \tag{B.8}$$

In all these cases, if (B.5) is verified, then $\delta_{.}(.) > 0$.

As explained above, the notations are used here to help the proof of efficiency of $f_{M,\tau}$. Thus, we use the following notations allowing simplification in the writings.

B.2.2 Characterization of the Sensor Set Over the Instants of Period τ

Definition 7. *We define the* ***characterization of the sensor set over the instants of period*** τ*:*

$$E_k := E(t_0 + k\tau)$$

In the same way, we define $(\Pi_k)_{k\in\mathbb{N}} := \{\Pi(t_0 + k\tau), k \in \mathbb{N}\}$.

For $i \in \Pi_k$*:*

$$E_{i,k} := E_i(t_0 + k\tau) = \left\{ \begin{array}{l} e_{i,k} = e_i(t_0 + k\tau) \\ p_{i,k} = p_i(t_0 + k\tau) \\ \delta_{i,k} = \delta_i(t_0 + k\tau) \end{array} \right.$$

.

From Definition 7, we mathematically define the **effectiveness** informally presented in Definition 3 (page 64):

Definition 8. *A period allocation function is said to be* ***effective over the instants***

of period τ *if, using characterization of the sensor set over the instants of period* τ*:*

$$\begin{aligned} \forall k \in \mathbb{N}, |\Pi_k| > 0 \Rightarrow \quad & \exists! i \in \Pi_k : \delta_{i,k} = \tau \\ & \forall j \in \Pi_k, j \neq i \Rightarrow \delta_{j,k} > \tau \end{aligned} \tag{B.9}$$

Then, its sample span L *is defined by:*

$$L := \max\{k, |\Pi_k| > 0\} + 1 \tag{B.10}$$

B.2.3 Preliminaries and Proof Scheme for the Effectiveness of $f_{M,\tau}$

First, we consider that the description of the algorithm is sufficient to assert that

Assertion 1. death-date *is a list updated at each new activation, such that it corresponds to the list sorted by ascending order of death of sensors whose relay is not already provided by other sensors.*

To prove the Proposition 2 (page 70), we propose a reasoning by induction. We prove the statement called $P(k)$:

Looking at the instant $t_0 + k\tau$:

- If the number of alive sensors does not exceed M, then the alive sensors activate exactly on the next consecutive $j\tau$ instants, $1 \leq j \leq |\Pi_k|$.
- Otherwise, the $j\tau$ instants are covered for the first M instants by M sensors. The next activation of the other sensors occurs at least $M\tau$ after the death of the next sensor. There is exactly one that activates $M\tau$ after the death of the next sensor, and all the rest activates after that time.

Mathematically, introducing the sensors not implied in the round-robin $\mathbb{I}_k = \{i \in \Pi_k, \delta_{i,k} > M\tau\}$, $P(k)$ can be mathematically written as:

$$\text{-If } |\Pi_k| \leq M, \forall j, 1 \leq j \leq |\Pi_k|, \exists! i \in \Pi_k : \delta_{i,k} = j\tau$$

$$\text{-Else,} \begin{cases} \forall j, 1 \leq j \leq M, \exists! i \in \Pi_k : \delta_{i,k} = j\tau \\ \exists! i \in \mathbb{I}_k : \delta_{i,k} = \text{next-death(k)} - (t_0 + k\tau) + M\tau \\ \forall i \in \mathbb{I}_k, \delta_{i,k} \geq \text{next-death(k)} - (t_0 + k\tau) + M\tau \end{cases}$$

Where next-death(k) represents the death time of the next sensor after $t_0 + k\tau$.

Thanks to the proof of $P(k)$, we will directly conclude that $f_{M,\tau}$ is effective on the time intervals of length τ based on Eq. (B.9).

B.2.4 Demonstration of Proposition 2 (Page 70)

Proof. Proof by induction:

Initialization: We want to prove $P(0)$. $|\Pi_0| = 1 \leq M$. $f_{M,\tau}$ sets the period to τ, so that $\delta_{0,0} = \tau$, hence $P(0)$ is true.

Heredity: Assume $P(k)$ is true for some $k \geq 0$. We define $n := |\Pi_k|$.

Disjunction of cases.

• **If $|\Pi_k| \leq M$ and $|\Pi_{k+1}| \leq M$.**

· If there is no variation in the set of alive sensors, $|\Pi_k| = |\Pi_{k+1}| = n \leq M$.

According to $P(k)$, sensors become active over the next j instances where $1 \leq j \leq n$. For sensors that remain inactive until a subsequent instance, say $1 < j$, there exists a unique sensor $i \in \Pi_k$ such that $\delta_{i,k} = j\tau$, which consequently implies $\delta_{i,k+1} = (j-1)\tau$.

For the sensor that activates at time $t_0 + (k+1)\tau$ (i such that $\delta_{i,k} = \tau$), according to the period update function Eq. (4.4), we have $f_{M,\tau} = |\Pi_{k+1}|\tau = n\tau$, which means $\delta_{i,k+1} = n\tau$.

Finally

$$\forall j, 1 \leq j \leq n, \exists! i : \delta_{i,k+1} = j\tau,$$

which means $P(k+1)$ is true.

· If a sensor with index l becomes inactive between states E_k and E_{k+1}, it means that $\delta_{l,k} = \tau$ and $\Pi_{k+1} = n-1$. Similar to the previous case, based on property $P(k)$, the other sensors activate in state E_k for the following instances where $1 < j \leq n$, which implies that in state E_{k+1} they activate at instances $1 \leq j < n$ (same reasoning as the previous point). Therefore, we directly observe that for $1 \leq j \leq n-1$, $\delta_{i,k} = j\tau$, leading to the conclusion that $P(k+1)$ holds true.

· We now consider a scenario in which m new sensors indexed as $(l+r)_{r\in[1,m]}$ initialize between states E_k and E_{k+1}, and no sensor becomes inactive; here, l represents the index of the last sensor that initialized before $t_0 + k\tau$.

Since $\Pi_{k+1} = \Pi_k \bigcup (l+r)_{r\in[1,m]} \leq M$, we have $n+m \leq M$. Without loss of generality, let us assume that they initialize in ascending order: $t_0 + k\tau < t_{l+1} < t_{l+2} ... < t_{l+m} \leq t_0 + (k+1)\tau$.

The period for all these new sensors is set using $f_{M,\tau}$:

$$f_{M,\tau}(\Pi(t_{l+r})) = \tau|\Pi(t_{l+r})| - (t_{l+r} - t_0)\%\tau$$

with $|\Pi(t_{l+r})| = n + r$, as per Eq. (4.4). Furthermore, it can be asserted that $t_{l+r} = t_0 + k\tau + (t_{l+r} - t_0)\%\tau$. Consequently, applying Eq. (B.8), we can expand $\delta_{l+r,k+1}$ as follows:

$$\begin{aligned}
&\forall r, 1 \leq r \leq m, \\
\delta_{l+r,k+1} &= t_{l+r} - (t_0 + (k+1)\tau) \\
&\quad + f_{M,\tau}(\Pi(t_{l+r})) \\
&= t_{l+r} - (t_0 + (k+1)\tau) \\
&\quad + (|\Pi(t_{l+r})|\tau - (t_{l+r} - t_0)\%\tau) \\
&= (|\Pi(t_{l+r})| - 1)\tau \\
&= (n + r - 1)\tau
\end{aligned}$$

Now from induction hypothesis, $\exists! i \in \Pi_k : \delta_{i,k} = \tau$, that thanks to $f_{M,\tau}$ implies that $\delta_{i,k+1} = (n+m)\tau$; plus, for $1 < j \leq n, \exists! i \in \Pi_k$ implying that $\delta_{i,k} = j\tau \Rightarrow \delta_{i,k+1} = (j-1)\tau$. Finally:

$$\forall j, 1 \leq j \leq n + m, \exists! i \in \Pi_{k+1} : \delta_{i,k} = j$$

Which means $P(k+1)$ is true.

It could also be possible to demonstrate that if between the states E_k and E_{k+1}, one sensor dies and several comes alive, $P(k+1)$ remains true.

• **If $|\Pi_{k+1}| > M$.**

· If there is no variation in the alive sensors $\Pi_k = \Pi_{k+1}$. From $P(k)$, $\forall j, 1 \leq j \leq M, \exists! i \in \Pi_k : \Pi_k = j\tau$. Then, since the period of these sensors is $M\tau$ (definition of $f_{M,\tau}$), from the same reasoning as above $\forall j, 1 \leq j \leq M, \exists! i \in \Pi_{k+1} : \Pi_{k+1} = j\tau$.

For the sensors which do not transmit on the first M instants, no sensor dies, so next-death(k) = next-death(k+1). From $P(k)$, $\exists! i \in \mathbb{I}_{k+1} : \delta_{i,k+1} =$ next-death(k+1) $-$ $(t_0 + (k+1)\tau) + M\tau$ and $\forall i \in \mathbb{I}_{k+1}, \delta_{i,k+1} \geq$ next-death(k+1) $- (t_0 + (k+1)\tau) + M\tau$. Hence, $P(k+1)$ is true.

· If m sensors, indexed as $(l+r)_{r\in[1,m]}$, become active and no sensors become inactive.

Firstly, if $|\Pi_k| < M$, then the initializing sensors follow the same pattern as a case studied previously, until $|\Pi_k| = M$. Without loss of generality, we now consider the situation where $(l+r)_{r\in[1,m]}$ become active while there are already M active sensors.

As seen previously, the sensor for which $\delta_{i,k} = \tau$ transitions to $\delta_{i,k+1} = M\tau$, and the

other sensors that were at instants $1 < j \leq M$ in state E_k shift to instants $1 \leq j < M$ in state E_{k+1}. This means that $\forall j, 1 \leq j \leq M, \exists! i \in \Pi_{k+1} : \delta_{i,k+1} = \tau j$.

Furthermore, the function $f_{M,\tau}$ sequentially sets the period of sensors indexed by $(l+r)r \in [1, m]$ to *death-date*$[0] - tl + 1 + M\tau$ and removes *death-date*$[0]$ from *death-date*. Consequently, this property holds for the sensors in the set $\mathbb{I}_{k+1}$, and thus $P(k+1)$ is true.

· If a sensor $l \in \Pi_k$ dies. Then, $\delta_{l,k} = \tau$ and $\Pi_{k+1} = \Pi_k / \{l\}$. Then next-death(k) $= t_0 + (k+1)\tau$. We consider, from hypothesis $P(k)$, i so that $\delta_{i,k} =$ next-death(k) $- (t_0 + k\tau) + M\tau = (M+1)\tau$ and $\delta_{i,k+1} = M\tau$. Hence :

$$\forall j, 1 \leq j \leq M, \exists! i \in \Pi_{k+1} : \Pi_{k+1} = j\tau$$

Moreover, next-death(k+1) is updated to the death date of the next sensor, and from assertion 1, and the definition of $f_{M,\tau}$, one sensor will transmit $M\tau$ after the death of the next sensor i.e. $\exists! i \in \mathbb{I}_{k+1} : \delta_{i,k+1} =$ next-death(k+1) $- (t_0 + (k+1)\tau) + M\tau$. Moreover, all the other sensors will transmit after that time: $\forall i \in \mathbb{I}_{k+1}, \delta_{i,k+1} \geq$ next-death(k+1) $- (t_0 + (k+1)\tau) + M\tau$. $P(k+1)$ is true.

Based on these base cases, it is possible to construct all possible cases to prove that if $P(k)$ is true, then $P(k+1)$ is also true.

Conclusion: For any $k \geq 0$, $P(k)$ holds true. This implies that $f_{M,\tau}$ is effective on the instants of period τ.

□

B.3 Bounds of the Sample Span of $f_{M,\tau}$

B.3.1 Preliminaries to the Proof of the Lower Bound

We aim to prove Eq. (4.5) (page 70). In reality, we develop our proof in two cases. By simplifying and assuming that the sensor consumes its entire energy, we can remove the floor functions in the expressions to obtain a single, easier-to-understand result.

One the one hand, for $\frac{n}{2} > M$:

$$\begin{array}{rcl} L(f_{M,\tau}) & \geq & \sum_{i=0}^{M-1} \left\lfloor \frac{e-c_e-(M-i+\mathbb{1}_{i>0})c_r}{c_e} \right\rfloor + (n-2M) * \left\lfloor \frac{e-c_e-2c_r}{c_e} \right\rfloor \\ & + & \sum_{i=n-M}^{n-1} \left\lfloor \frac{e-c_e-(i-n+M+2)c_r}{c_e} \right\rfloor \\ & = & \sum_{i=0}^{M-1} \frac{e-c_e-(i+1+\mathbb{1}_{i>0})c_r}{c_e} + (n-2M) * \frac{e-c_e-2c_r}{c_e} \\ & + & \sum_{i=0}^{M-1} \frac{e-c_e-(i+2)c_r}{c_e} \\ & = & \sum_{i=0}^{M-1} \frac{2e-2c_e-(2i+3+\mathbb{1}_{i>0})c_r}{c_e} + (n-2M) * \frac{e-c_e-2c_r}{c_e} \\ & = & \frac{n(2e-2c_e)-(4M-1)c_r}{c_e} + \frac{M(M-1)}{c_e} + (n-2M) * \frac{e-c_e-2c_r}{c_e} \\ & = & \frac{ne-nc_e-(2n-1+M(M-1))c_r}{c_e} \end{array} \quad \text{(B.11)}$$

Furthermore, for $M \geq \frac{n}{2}$, we have another formulae. Observing that $2M - n + \mathbb{1}_{i>0} = (M - i + \mathbb{1}_{i>0}) + (i - n + M)$, we have that:

$$\begin{array}{rcl} L(f_{M,\tau}) & \geq & \sum_{i=0}^{n-M-1} \left\lfloor \frac{e-c_e-(M-i+\mathbb{1}_{i>0})c_r}{c_e} \right\rfloor \\ & + & \sum_{i=n-M}^{M-1} \left\lfloor \frac{e-c_e-(2M-n+\mathbb{1}_{i>0})c_r}{c_e} \right\rfloor \\ & + & \sum_{i=M}^{n-1} \left\lfloor \frac{e-c_e-(i-n+M+2)c_r}{c_e} \right\rfloor \\ & = & \frac{ne-nc_e}{c_e} - \left(\sum_{i=0}^{M-1} \frac{M-i+\mathbb{1}_{i>0}}{c_e} + \sum_{i=n-M}^{n-1} \frac{i-n+M}{c_e} + \sum_{i=M}^{n-1} \frac{2}{c_e} \right) c_r \\ & = & \frac{ne-nc_e}{c_e} - \left(\frac{M(M+1)}{2c_e} + \frac{M-1}{c_e} + \frac{M(M-1)}{2c_e} + \frac{(n-M)*2}{c_e} \right) c_r \\ & = & \frac{ne-nc_e}{c_e} - \left(\frac{M(M-1)}{c_e} + \frac{M+(M-1)+2n-2M}{c_e} \right) c_r \\ & = & \frac{ne-nc_e-(2n-1+M(M-1))c_r}{c_e} \end{array} \quad \text{(B.12)}$$

B.3.2 Demonstration of the Lower Bound of $f_{M,\tau}$

Proof. Let a sensor indexed i. It modifies $1 + \mathbb{1}_{i>0}$ times its activation period to adjust itself with respect to the other sensors during its first two activations.

$|\Pi(.)|$ varies at each initialization and death of a sensor. We will study the case where a sensor i is subject to the most period changes.

· If i comes alive while $|\Pi(.)| \leq M$, it can need to modify its activation period for any new sensor activating, until there are M alive sensors i.e. at most $M - i - 1$ additional times.

In that case, the sensor i needs to transmit at least once between each initialization. This is the case if the time between two initializations is greater than the period of i. We can quantify a sufficient condition to have this realized:

$$\forall i, 1 \leq i \leq M-1, t_i - t_{i-1} > \tau i$$

· Furthermore, if we consider a sensor i that is alive when there are no more than M remaining sensors alive, it could need to change its activation period for each sensor death from the time when there are M remaining sensors. Considering that the sensors are alive at sufficiently spaced instants, and that they have the same initial energy, then the initialization index corresponds to the death index of the sensors. Thus, the sensor i will change its activation period at the death of sensors if $i \geq n - M$. It will thus change at most $i - (n - M)$ additional times.

· In the final case, the sensor takes over the responsibilities of a sensor that has just died, and it maintains its activation period at $M\tau$ thereafter. This results in two period changes.

Depending on the chosen value of M, they are 2 cases:

•If M is small relative to n i.e. $M < \frac{n}{2}$.

The sensors will have a different number of period changes following these 3 intervals, as also illustrated in Fig. B.1:

$$\begin{aligned} &\text{If } 0 \leq i \leq M-1, && r_i = M - i + \mathbb{1}_{i>0} \\ &\text{If } M-1 \leq i \leq n-M, && r_i = 2 \\ &\text{If } n-M \leq i \leq n-1, && r_i = i - (n-M) + 2 \end{aligned}$$

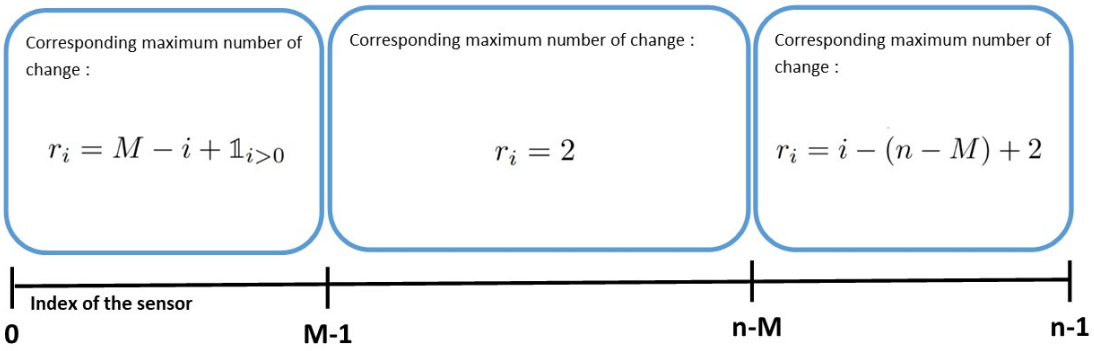


Figure B.1: Representation of the number of period changes depending on the index of the sensor when $M < \frac{n}{2}$

We can define the number of activations on the time steps τ without duplicates, and

thus have a lower bound formula for the sample span:

$$\begin{aligned} L(f_{M,\tau}) \geq & \quad \sum_{i=0}^{M-1} \left\lfloor \frac{e-c_e-(M-i+\mathbb{1}_{i>0})c_r}{c_e} \right\rfloor \\ + & \quad (n-2M) * \left\lfloor \frac{e-c_e-2c_r}{c_e} \right\rfloor \\ + & \quad \sum_{i=n-M}^{n-1} \left\lfloor \frac{e-c_e-(i-n+M+2)c_r}{c_e} \right\rfloor \end{aligned}$$

•If M is close to n i.e. $M \geq \frac{n}{2}$.

In this case, the intervals $[0, M-1]$ and $[n-M, n-1]$ overlap. We study the 3 following intervals (Fig. B.2):

$$\begin{aligned} &\text{If } 0 \leq i \leq n-M, && r_i = M - i + \mathbb{1}_{i>0} \\ &\text{If } n-M \leq i \leq M-1, && r_i = 2M - n + \mathbb{1}_{i>0} \\ &\text{If } M-1 \leq i \leq n-1, && r_i = i - (n-M) + 2 \end{aligned}$$

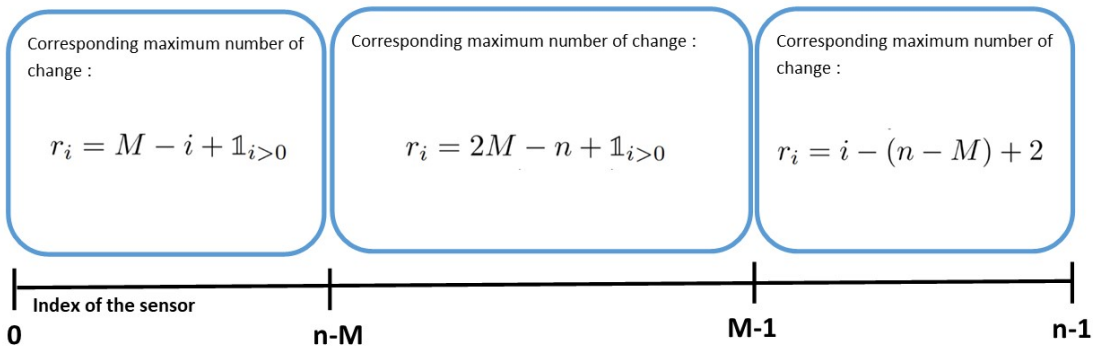


Figure B.2: Representation of the number of period changements depending on the index of the sensor when $M \geq \frac{n}{2}$

We can then get the minimization:

$$\begin{aligned} L(f_{M,\tau}) \geq & \quad \sum_{i=0}^{n-M-1} \left\lfloor \frac{e-c_e-(M-i+\mathbb{1}_{i>0})c_r}{c_e} \right\rfloor \\ + & \quad \sum_{i=n-M}^{M-1} \left\lfloor \frac{e-c_e-(2M-n+\mathbb{1}_{i>0})c_r}{c_e} \right\rfloor \\ + & \quad \sum_{i=M}^{n-1} \left\lfloor \frac{e-c_e-(i-n+M+2)c_r}{c_e} \right\rfloor \end{aligned}$$

□

This leads us directly to the simplified expression Eq. (4.6) (page 70).

B.3.3 Demonstration of the Upper Bound of $f_{M,\tau}$

We will here demonstrate:

$$L(f_{M,\tau}) \leq \sum_{i=0}^{n-1} \left\lfloor \frac{e - c_e - c_r(1 + \mathbb{1}_{M>1 \text{ or } i>0})}{c_e} \right\rfloor$$

that will directly lead us, by inequality computations, to the more understandable expression without the integer parts in Eq. (4.6).

Proof. We consider that the first M sensors come alive in the first time interval τ i.e.

$$\forall i, 1 \leq i \leq M, t_i \in [t_0, t_0 + \tau]$$

Let us consider these first M sensors. The first sensor of index 0 : $f_{M,\tau}$ modifies its activation period a first time to τ, then modifies it to $M\tau$ if $M \neq 1$. For the sensors of index $0 < i < M$, their first activation period is $\tau|\Pi(t_i)| - (t - t_0)\%\tau$. Their second period is exactly τM. Each sensor performs exactly $r_i = 2$ period changes.

Moreover, for the following sensors, they also change their activation period a first time to transmit following the death of a sensor, then the period is fixed to $M\tau$. Finally, since all the sensors turned on at the same time and consumed a similar amount of energy, if $n \equiv 0[M]$, then the cycles of M sensors will be renewed each time at the same time, and thus the M last sensors will die at the same time in turn (no additional period change consumption). In this case, all sensors must change their activation period twice, except for the 0 sensor, if $M = 1$:

$$L(f_{M,\tau}) \leq \sum_{i=0}^{n-1} \left\lfloor \frac{e - c_e - c_r(1 + \mathbb{1}_{M>1 \text{ or } i>0})}{c_e} \right\rfloor$$

□

Appendix C

Proofs of Diversity Properties of Synchronous Round-Robin in Chapter 5

In this part of appendix, we provide proof of the remarkable properties of the period allocation function Synchronous Round-Robin. Specifically, we demonstrate that for a global reception frequency, it is necessary for all sensor activation periods to be equal in Appendix C.1. Additionally, starting with a set of sensors transmitting with activation periods of $n\tau$, where the allocation function is effective (i.e., sensors transmit over the instants of period τ), we show that if a sensor is shifted away from its interval, the minimum diversity is reduced, in Appendix C.2.

C.1 Same Activation Periods to Maximize the Mean Diversity

Let us consider a set of sensors Π, sending periodic messages with constant periods $(p_i)_{i\in\Pi}$.

We look at the solutions where the mean number of message per time unit is fixed to $\frac{1}{\tau}$:

$$\sum_{i\in\Pi} \frac{1}{p_i} = \frac{1}{\tau}$$

By definition, a freshness function is a positive decreasing function. The proof holds for a strictly monotonic and differentiable function $f(x)$, with the derivative denoted as $f'(x)$ and a primitive function denoted as $F(x)$.

The mean diversity is the sum of the mean freshness of all sensors. As the sensors activate periodically with a constant period, we can compute the mean freshness over one

period. For the sensor i with period p_i, it freshness F_i is:

$$\begin{aligned} F_i &= \frac{1}{p_i}\int_0^{p_i} f(t)dt \\ &= \frac{F(p_i)-F(0)}{p_i} \end{aligned}$$

with the mean diversity $D = \sum_{i\in\Pi} F_i$.

This brings us to the search for:

$$\begin{aligned} i \in \Pi, 0 < p_i, \quad & max(\textstyle\sum_{i\in\Pi} \frac{F(p_i)-F(0)}{p_i}) \\ & \textstyle\sum_{i\in\Pi} \frac{1}{p_i} = \frac{1}{\tau} \end{aligned}$$

We then introduce the Lagrangian:

$$L((p_i)_{i\in\Pi}, \lambda) = \sum_{i\in\Pi} \frac{F(p_i) - F(0)}{p_i} + (\sum_{i\in\Pi} \frac{1}{p_i} - \frac{1}{\tau})\lambda$$

Since we are looking for an extremum, all partial derivatives are null:

$$\begin{aligned} \forall j \in \Pi, \quad & \frac{\delta L}{\delta p_j} = \frac{f(p_j)p_j - (F(p_j)-F(0))}{p_j^2} - \frac{\lambda}{p_j^2} = 0 \\ & \frac{\delta L}{\delta \lambda} = \textstyle\sum_{i\in\Pi} \frac{1}{p_i} - \frac{1}{\tau} = 0 \end{aligned}$$

Let us develop the first equation:

$$\begin{aligned} \frac{\delta L}{\delta p_j} = 0 &\Leftrightarrow \frac{f(p_j)p_j - (F(p_j)-F(0)) - \lambda}{p_j^2} = 0 \\ &\Leftrightarrow \lambda = f(p_j)p_j - (F(p_j) - F(0)) \end{aligned}$$

We define $g(x) = f(x)x - (F(x) - F(0))$, so that:

$$\begin{aligned} g'(x) &= f'(x)x + f(x) - f(x) \\ &= f'(x)x \end{aligned}$$

Since f is strictly monotonous, f' keeps the same sign, so does $g'(x)$ for $x > 0$. This indicates that g is strictly monotonic. Consequently, $g(x) = \lambda$ has a unique pre-image, implying that:

$$\forall j \in \Pi, \lambda = f(p_j)p_j - (F(p_j) - F(0)) \implies p_j = g^{-1}(\lambda)$$

All p_j are the same, and equal to $|\Pi|\tau$ with $|\Pi|$ the number of sensors.

The mean diversity is then $D = \frac{F(|\Pi|\tau)-F(0)}{\tau}$

C.2 Regular Time Stamps to Maximize Minimum Diversity

We consider a set of sensors indexed in increasing order as $\Pi = [0, n-1]$, all transmitting at a period of $n\tau$. Since these sensors activate at the same period, our analysis can be confined to a segment of size $n\tau$. By assuming that the activation of sensor 0 occurs at $t = 0$ without any loss of generality, we can effectively focus our study on the segment $[0, n\tau[$. Thus, we consider $t_0 = 0 < t_1 ... < t_{n-1} < n\tau$.

Let's consider f as a strictly decreasing function.

The freshness of sensor $i \in [0, n-1]$ transmitting at t_i over time is defined as:

$$\begin{aligned} f_i(t) &= f(n\tau - t_i + t) && \text{if } t < t_i \\ &= f(t - t_i) && \text{if } t \geq t_i \end{aligned}$$

Specifically, the global minimum diversity occurs at one of the local minima located at t_j^-, as the diversity function is decreasing on the intervals $]t_{j-1}, t_j[$. The diversity at these points is:

$$\begin{aligned} D(0^-) &= \textstyle\sum_{i=0}^{n-1} f(n\tau - t_i) \\ D(t_j^-) &= \textstyle\sum_{i=0}^{j-1} f(t_j - t_i) + \sum_{i=j}^{n-1} f(n\tau - t_i + t_j) \end{aligned} \tag{C.1}$$

By taking $t_j = j\tau$, we obtain:

$$\begin{aligned} D(0^-) &= \textstyle\sum_{i=0}^{n-1} f((n-i)\tau) \\ &= \textstyle\sum_{i=1}^{n} f(i\tau) \\ D(t_j^-) &= \textstyle\sum_{i=0}^{j-1} f((j-i)\tau) + \sum_{i=j}^{n-1} f((n-j+i)\tau) \\ &= \textstyle\sum_{i=1}^{j} f(i\tau) + \sum_{i=j+1}^{n} f(i\tau) \\ &= \textstyle\sum_{i=1}^{n} f(i\tau) \end{aligned}$$

In particular, $D(0^-) = ... = D(t_{n-1}^-) = \sum_{i=1}^{n} f(i\tau) = \min(D(t))$. The scheduling of activations for 5 sensors is illustrated in Fig. C.1.

Now, let's consider moving one of the activations, say t_j, from its original scheduling at $j\tau$. Without loss of generality, for $j \in [1, n-2]$, let's examine the case where t_j is

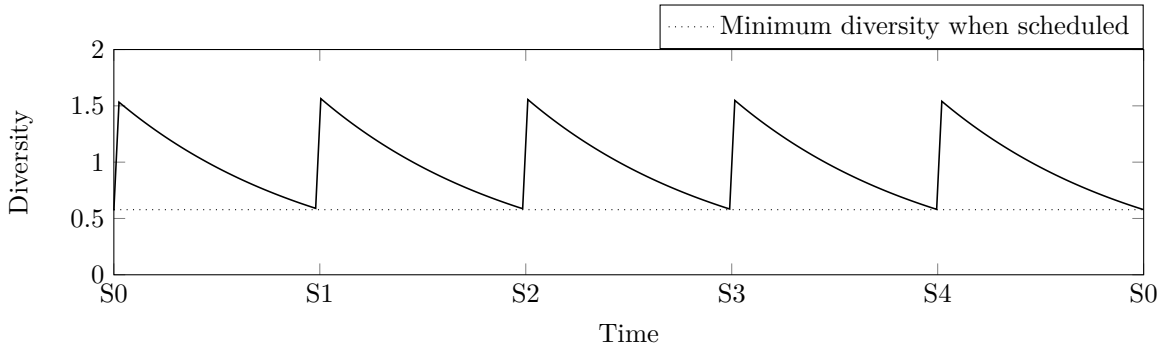


Figure C.1: Diversity over time when 5 sensors have the same activation period and are scheduled to receive observations at regular intervals. Time is represented on the x-axis, with points S0, S1, S2, S3, and S4 indicating sensor activations; diversity is shown on the y-axis.

moved, i.e., $t_j \neq j\tau$.

• If we move it to the right: $j\tau < t_j < (j+1)\tau$, then for all i where $i \neq j$, we have $f(t_j - i\tau) < f(j\tau - i\tau)$ and $f(n\tau + t_j - i\tau) < f(n\tau + j\tau - i\tau)$ due to the strict decreasing nature of f. By evaluating the diversity at t_j^-, using Eq. (C.1):

$$\begin{aligned} D(t_j^-) &= \textstyle\sum_{i=0}^{j-1} f(t_j - i\tau) + \sum_{i=j}^{n-1} f(n\tau + t_j - i\tau) \\ &< \textstyle\sum_{i=0}^{j-1} f(j\tau - i\tau) + \sum_{i=j}^{n-1} f(n\tau + j\tau - i\tau) \\ &= \textstyle\sum_{i=1}^{n} f(i\tau) \end{aligned}$$

By evaluating the diversity at t_j^-, we observe that the diversity is lower, indicating that the overall global minimum diversity is lower when we move t_j from its original scheduling at $j\tau$. In Fig. C.2, we illustrate the reduction of the minimum diversity by moving the activation of S3 to the right.

• If t_j is moved to the left: $t_j < j\tau < (i-1)\tau$, then $f((j+1)\tau - t_j) < f((j+1)\tau - j\tau) = f(\tau)$. By evaluating the diversity at $t_{j+1}^- = (j+1)\tau$:

$$\begin{aligned} f(t_{j+1}) &= f(t_{j+1}) + \textstyle\sum_{i=0}^{j} f(t_{j+1} - t_i) + \sum_{i=j+1}^{n-1} f(n\tau + t_j - t_i) \\ &= \textstyle\sum_{i=2}^{n} f(i\tau) + f((j+1)\tau - t_j) \\ &< \textstyle\sum_{i=1}^{n} f(i\tau) \end{aligned}$$

We also observe that by moving t_j to the left, the minimum diversity becomes lower, as illustrated in Fig. C.3.

Overall, by locally moving an observation t_j away from its scheduled time $j\tau$, the

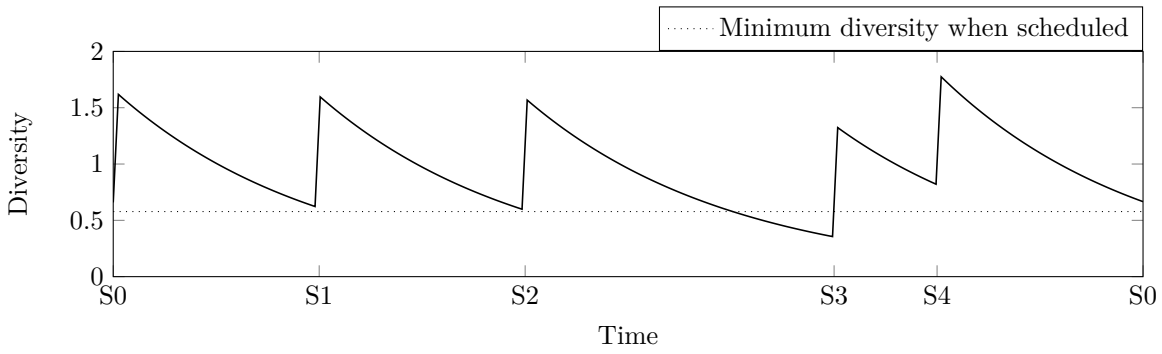


Figure C.2: Diversity when sensors have the same activation period, moving $S3$ to the right.

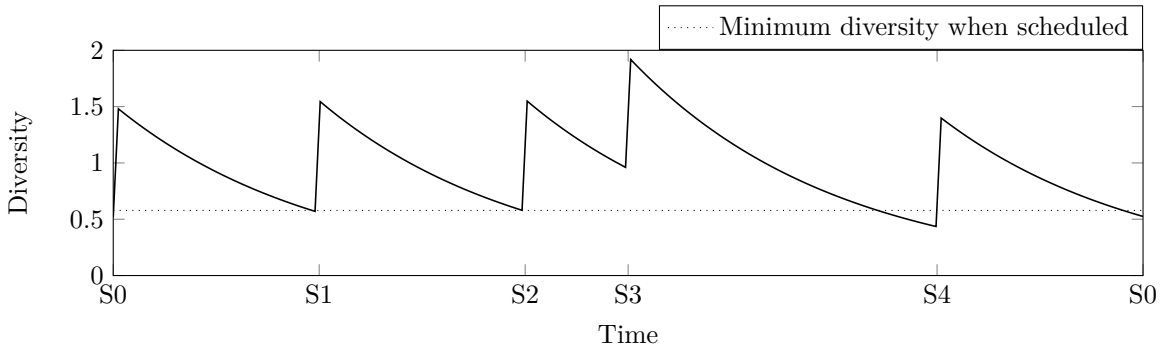


Figure C.3: Diversity among sensors with the same activation period, moving $S3$ to the left.

minimum diversity is reduced.

Intuitively, arranging the sensors in a strict order would likely maximize the overall minimum diversity, although we have not found a formal proof for this.

Appendix D

Calculations for Simple Kriging Resolution

From the definition of $\hat{\theta}_{\hat{t}}$, we can already establish through its expectation calculation that it is unbiased: $E[\hat{\theta}_{\hat{t}}] = E[\Theta_t] = 0$. Furthermore, by expanding their squared difference defined in Eq. (6.2), we have:

$$\begin{aligned}
\Delta(\hat{t}) &= E[(\psi_{\hat{t}}^\top \theta - \Theta_{\hat{t}})^2] \\
&= E[\psi_{\hat{t}}^\top \theta\theta^\top \psi_{\hat{t}} - \Theta_{\hat{t}}\theta^\top \psi_{\hat{t}} - \psi_{\hat{t}}^\top \theta\Theta_{\hat{t}} + \Theta_{\hat{t}}^2] \\
&= \psi_{\hat{t}}^\top E[\theta\theta^\top]\psi_{\hat{t}} - 2E[\Theta_{\hat{t}}\theta^\top]\psi_{\hat{t}} + E[\Theta_{\hat{t}}]^2 \\
&= \psi_{\hat{t}}^\top K\psi_{\hat{t}} - 2k_{\hat{t}}^\top \psi_{\hat{t}} + \sigma_{\hat{t}}^2
\end{aligned}$$

Where $\sigma_{\hat{t}} = E[\Theta_{\hat{t}}]$, independent of $\psi_{\hat{t}}$, leading to Eq. (6.3).

We aim to find the vector $\psi_{\hat{t}}$ that minimizes $\Delta(\hat{t})$. The derivative with respect to each $\psi_{t,\hat{t}}$ is zero, resulting in:

$$\begin{aligned}
&\frac{\partial\Delta(\hat{t})}{\partial\psi_{\hat{t}}} = 2K\psi_{\hat{t}} - 2k_{\hat{t}} = 0 \\
\Leftrightarrow\quad &\psi_{\hat{t}} = K^{-1}k_{\hat{t}}
\end{aligned}$$

K is a symmetric matrix, so K^{-1} is a symmetric matrix, leading us to the expression of the estimation $\hat{\theta}_{\hat{t}}$ defined in Eq. (6.4):

$$\hat{\theta}_{\hat{t}} = k_{\hat{t}}^\top K^{-1}\theta$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[AA20] Ali Kadhum M. Al-Qurabat and Suha Abdulhussein Abdulzahra, "An Overview of Periodic Wireless Sensor Networks to The Internet of Things", en, *in*: *IOP Conference Series: Materials Science and Engineering* 928.*3* (Nov. 2020), Publisher: IOP Publishing, p. 032055, ISSN: 1757-899X, DOI: 10.1088/1757-899X/928/3/032055, URL: https://dx.doi.org/10.1088/1757-899X/928/3/032055 (visited on 04/17/2023).

[AC16] Ozgur Umut Akgul and Berk Canberk, "Self-Organized Things (SoT): An energy efficient next generation network management", en, *in*: *Computer Communications* 74 (Jan. 2016), pp. 52–62, ISSN: 01403664, DOI: 10.1016/j.comcom.2014.07.004, URL: https://linkinghub.elsevier.com/retrieve/pii/S0140366414002473 (visited on 11/28/2022).

[AC17] Samaneh Aminikhanghahi and Diane J. Cook, "A survey of methods for time series change point detection", en, *in*: *Knowledge and Information Systems* 51.*2* (May 2017), pp. 339–367, ISSN: 0219-1377, 0219-3116, DOI: 10.1007/s10115-016-0987-z, URL: http://link.springer.com/10.1007/s10115-016-0987-z (visited on 12/01/2022).

[Ade+17] Ferran Adelantado et al., "Understanding the Limits of LoRaWAN", en, *in*: *IEEE Communications Magazine* 55.*9* (2017), pp. 34–40, ISSN: 0163-6804, DOI: 10.1109/MCOM.2017.1600613, URL: http://ieeexplore.ieee.org/document/8030482/ (visited on 12/02/2022).

[Aky+02] I. F. Akyildiz et al., "Wireless sensor networks: a survey", en, *in*: *Computer Networks* 38.*4* (Mar. 2002), pp. 393–422, ISSN: 1389-1286, DOI: 10.1016/S1389-1286(01)00302-4, URL: https://www.sciencedirect.com/science/article/pii/S1389128601003024 (visited on 02/23/2023).

[Ali+09] C. Alippi et al., "Energy management in wireless sensor networks with energy-hungry sensors", en, *in*: *IEEE Instrumentation & Measurement Magazine* 12.*2* (Apr. 2009), pp. 16–23, ISSN: 1094-6969, DOI: 10.1109/MIM.2009.

4811133, URL: http://ieeexplore.ieee.org/document/4811133/ (visited on 11/28/2022).

[Alm+16] Fernando R. Almeida et al., "Fractal Clustering and similarity measure: Two new approaches for reducing energy consumption in Wireless Sensor Networks", *in*: *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, ISSN: 2165-8536, July 2016, pp. 288–293, DOI: 10.1109/ICUFN.2016.7537034.

[Ana+09] Giuseppe Anastasi et al., "Energy conservation in wireless sensor networks: A survey", en, *in*: *Ad Hoc Networks* 7.*3* (May 2009), pp. 537–568, ISSN: 15708705, DOI: 10.1016/j.adhoc.2008.06.003, URL: https://linkinghub.elsevier.com/retrieve/pii/S1570870508000954 (visited on 12/01/2022).

[APM09] Bakhtiar Qutub Ali, Niki Pissinou, and Kia Makki, "Identification and Validation of Spatio-Temporal Associations in Wireless Sensor Networks", en, *in*: *2009 Third International Conference on Sensor Technologies and Applications*, Athens, Greece: IEEE, June 2009, pp. 496–501, ISBN: 978-0-7695-3669-9, DOI: 10.1109/SENSORCOMM.2009.83, URL: http://ieeexplore.ieee.org/document/5210870/ (visited on 12/01/2022).

[ASY15] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah, "Time-series clustering – A decade review", en, *in*: *Information Systems* 53 (Oct. 2015), pp. 16–38, ISSN: 03064379, DOI: 10.1016/j.is.2015.04.007, URL: https://linkinghub.elsevier.com/retrieve/pii/S0306437915000733 (visited on 07/18/2023).

[Azi+13] A. A. Aziz et al., "A Survey on Distributed Topology Control Techniques for Extending the Lifetime of Battery Powered Wireless Sensor Networks", en, *in*: *IEEE Communications Surveys & Tutorials* 15.*1* (2013), pp. 121–144, ISSN: 1553-877X, DOI: 10.1109/SURV.2012.031612.00124, URL: http://ieeexplore.ieee.org/document/6177190/ (visited on 11/28/2022).

[Bah+14] Jacques Bahi et al., "Efficient distributed lifetime optimization algorithm for sensor networks", en, *in*: *Ad Hoc Networks* 16 (May 2014), pp. 1–12, ISSN: 15708705, DOI: 10.1016/j.adhoc.2013.11.010, URL: https://linkinghub.elsevier.com/retrieve/pii/S1570870513002795 (visited on 11/29/2022).

[Ban+22] Konstantina Banti et al., "LoRaWAN Communication Protocols: A Comprehensive Survey under an Energy Efficiency Perspective", en, *in*: *Telecom* 3.*2* (May 2022), pp. 322–357, ISSN: 2673-4001, DOI: 10.3390/telecom3020018, URL: https://www.mdpi.com/2673-4001/3/2/18 (visited on 11/30/2022).

[BDK16] Alexandros-Apostolos A. Boulogeorgos, Panagiotis D. Diamantoulakis, and George K. Karagiannidis, *Low Power Wide Area Networks (LPWANs) for Internet of Things (IoT) Applications: Research Challenges and Future Trends*, en, arXiv:1611.07449 [cs], Nov. 2016, URL: http://arxiv.org/abs/1611.07449 (visited on 12/02/2022).

[Bou+18] Taoufik Bouguera et al., "Energy consumption modeling for communicating sensors using LoRa technology", en, *in*: *2018 IEEE Conference on Antenna Measurements & Applications (CAMA)*, Västerås: IEEE, Sept. 2018, pp. 1–4, ISBN: 978-1-5386-5795-9, DOI: 10.1109/CAMA.2018.8530593, URL: https://ieeexplore.ieee.org/document/8530593/ (visited on 12/02/2022).

[Bou04] Mokrane Bouzeghoub, "A framework for analysis of data freshness", en, *in*: *Proceedings of the 2004 international workshop on Information quality in information systems*, Paris France: ACM, June 2004, pp. 59–67, ISBN: 978-1-58113-902-0, DOI: 10.1145/1012453.1012464, URL: https://dl.acm.org/doi/10.1145/1012453.1012464 (visited on 06/08/2023).

[Bou07] Mounzer Boubou, *Contribution aux méthodes de classification non supervisée via des approches prétopologiques et d'agrégation d'opinions*, fr, 2007.

[Bur+09] Chiara Buratti et al., "An Overview on Wireless Sensor Networks Technology and Evolution", en, *in*: *Sensors* 9.*9* (Aug. 2009), pp. 6869–6896, ISSN: 1424-8220, DOI: 10.3390/s90906869, URL: http://www.mdpi.com/1424-8220/9/9/6869 (visited on 04/14/2023).

[Car+02] Mihaela Cardei et al., "Wireless Sensor Networks with Energy Efficient Organization", en, *in*: *Journal of Interconnection Networks* 03.*03n04* (Sept. 2002), pp. 213–229, ISSN: 0219-2659, 1793-6713, DOI: 10.1142/S021926590200063X, URL: https://www.worldscientific.com/doi/abs/10.1142/S021926590200063X (visited on 11/28/2022).

[Cas+10] Charles C. Castello et al., "Optimal sensor placement strategy for environmental monitoring using Wireless Sensor Networks", en, *in*: *2010 42nd Southeastern Symposium on System Theory (SSST 2010)*, Tyler, TX, USA: IEEE, Mar. 2010, pp. 275–279, ISBN: 978-1-4244-5690-1, DOI: 10.1109/SSST.2010.5442825, URL: http://ieeexplore.ieee.org/document/5442825/ (visited on 11/28/2022).

[CH20] Sultan Mahmood Chowdhury and Ashraf Hossain, "Different Energy Saving Schemes in Wireless Sensor Networks: A Survey", en, *in*: *Wireless Personal Communications* 114.*3* (Oct. 2020), pp. 2043–2062, ISSN: 0929-6212, 1572-834X, DOI: 10.1007/s11277-020-07461-5, URL: https://link.springer.com/10.1007/s11277-020-07461-5 (visited on 11/30/2022).

[Cha09] Varun Chandola, "Anomaly Detection : A Survey", en, *in*: *ACM Computing Surveys* (2009), p. 72.

[Che+12] Qian Chen et al., "Learning optimal warping window size of DTW for time series classification", *in*: *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, July 2012, pp. 1272–1277, DOI: 10.1109/ISSPA.2012.6310488.

[Chu+15] Monika Chuchro et al., "A Concept of Time Windows Length Selection in Stream Databases in the Context of Sensor Networks Monitoring", en, *in*: *New Trends in Database and Information Systems II*, ed. by Nick Bassiliades et al., Advances in Intelligent Systems and Computing, Cham: Springer International Publishing, 2015, pp. 173–183, ISBN: 978-3-319-10518-5, DOI: 10.1007/978-3-319-10518-5_14.

[Chu95] Chia-Shang James Chu, "Time series segmentation: A sliding window approach", en, *in*: *Information Sciences* 85.*1* (July 1995), pp. 147–173, ISSN: 0020-0255, DOI: 10.1016/0020-0255(95)00021-G, URL: https://www.sciencedirect.com/science/article/pii/002002559500021G (visited on 04/24/2023).

[CKJ05] Chong Liu, Kui Wu, and Jian Pei, "A dynamic clustering and scheduling approach to energy saving in data collection from wireless sensor networks", en, *in*: *2005 Second Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2005. IEEE SECON 2005.* Santa Clara, CA, USA: IEEE, 2005, pp. 374–385, ISBN: 978-0-7803-

9011-9, DOI: 10.1109/SAHCN.2005.1557091, URL: http://ieeexplore.ieee.org/document/1557091/ (visited on 04/13/2023).

[CN16] Jui-Sheng Chou and Ngoc-Tri Ngo, "Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns", en, *in*: *Applied Energy* 177 (Sept. 2016), pp. 751–770, ISSN: 0306-2619, DOI: 10.1016/j.apenergy.2016.05.074, URL: https://www.sciencedirect.com/science/article/pii/S0306261916306717 (visited on 04/24/2023).

[Cor+07] Luiz H.A. Correia et al., "Transmission power control techniques for wireless sensor networks", en, *in*: *Computer Networks* 51.*17* (Dec. 2007), pp. 4765–4779, ISSN: 13891286, DOI: 10.1016/j.comnet.2007.07.008, URL: https://linkinghub.elsevier.com/retrieve/pii/S1389128607002034 (visited on 11/30/2022).

[CQ98] C. Caruso and F. Quarta, "Interpolation methods comparison", en, *in*: *Computers & Mathematics with Applications* 35.*12* (June 1998), pp. 109–126, ISSN: 08981221, DOI: 10.1016/S0898-1221(98)00101-1, URL: https://linkinghub.elsevier.com/retrieve/pii/S0898122198001011 (visited on 07/11/2023).

[Cre93] Cressie Noel A C, *Statistics for spatial data / Noel A. C. Cressie,...* eng, Revised edition, Wiley series in probability and mathematical statistics Applied probability and statistics, New York [etc: John Wiley & Sons, Inc, 1993, ISBN: 978-0-471-00255-0.

[CS16] Robert M. Curry and J. Cole Smith, "A survey of optimization algorithms for wireless sensor network lifetime maximization", en, *in*: *Computers & Industrial Engineering* 101 (Nov. 2016), pp. 145–166, ISSN: 03608352, DOI: 10.1016/j.cie.2016.08.028, URL: https://linkinghub.elsevier.com/retrieve/pii/S0360835216303321 (visited on 11/30/2022).

[CW06] Sesh Commuri and Mohamed K. Watfa, "Coverage Strategies in Wireless Sensor Networks", en, *in*: *International Journal of Distributed Sensor Networks* 2.*4* (Oct. 2006), pp. 333–353, ISSN: 1550-1477, 1550-1477, DOI: 10.1080/15501320600719151, URL: http://journals.sagepub.com/doi/10.1080/15501320600719151 (visited on 11/29/2022).

[DB12] F. De Rango and E. Bragina, "A weighted spatial correlation based strategy to improve the event estimation in a Wireless Sensor Networks", en, *in*: *2012 IFIP Wireless Days*, Dublin, Ireland: IEEE, Nov. 2012, pp. 1–6, ISBN: 978-1-4673-4404-3 978-1-4673-4402-9 978-1-4673-4403-6, DOI: 10.1109/WD.2012.6402884, URL: http://ieeexplore.ieee.org/document/6402884/ (visited on 12/01/2022).

[DBO17] Gabriel Martins Dias, Boris Bellalta, and Simon Oechsner, "A Survey About Prediction-Based Data Reduction in Wireless Sensor Networks", en, *in*: *ACM Computing Surveys* 49.*3* (Sept. 2017), pp. 1–35, ISSN: 0360-0300, 1557-7341, DOI: 10.1145/2996356, URL: https://dl.acm.org/doi/10.1145/2996356 (visited on 11/30/2022).

[Déj+07] S. Déjean et al., "Clustering Time-Series Gene Expression Data Using Smoothing Spline Derivatives", en, *in*: *EURASIP Journal on Bioinformatics and Systems Biology* 2007 (2007), pp. 1–10, ISSN: 1687-4145, DOI: 10.1155/2007/70561, URL: http://bsb.eurasipjournals.com/content/2007/1/70561 (visited on 07/18/2023).

[DHT19] Valentina Di Vincenzo, Martin Heusse, and Bernard Tourancheau, "Improving Downlink Scalability in LoRaWAN", en, *in*: *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China: IEEE, May 2019, pp. 1–7, ISBN: 978-1-5386-8088-9, DOI: 10.1109/ICC.2019.8761157, URL: https://ieeexplore.ieee.org/document/8761157/ (visited on 12/02/2022).

[DKP19] Ioannis Daramouskas, Vaggelis Kapoulas, and Theodoros Pegiazis, "A survey of methods for location estimation on Low Power Wide Area Networks", en, *in*: *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, PATRAS, Greece: IEEE, July 2019, pp. 1–4, ISBN: 978-1-72814-959-2, DOI: 10.1109/IISA.2019.8900701, URL: https://ieeexplore.ieee.org/document/8900701/ (visited on 11/28/2022).

[Dur01] Nicolas Durrande, "Étude de classes de noyaux adaptées à la simplification et à l'interprétation des modèles d'approximation. Une approche fonctionnelle et probabiliste.", fr, PhD thesis, {Ecole Nationale Sup{\'e}rieure des Mines de Saint-Etienne}, 2001, URL: https://theses.hal.science/tel-00770625.

[Elh+17] Mostafa Elhoushi et al., "A Survey on Approaches of Motion Mode Recognition Using Sensors", en, *in*: *IEEE Transactions on Intelligent Transportation Systems* 18.*7* (July 2017), pp. 1662–1686, ISSN: 1524-9050, 1558-0016, DOI: 10.1109/TITS.2016.2617200, URL: http://ieeexplore.ieee.org/document/7726001/ (visited on 04/18/2023).

[Eng+18] Felicia Engmann et al., "Prolonging the Lifetime of Wireless Sensor Networks: A Review of Current Techniques", en, *in*: *Wireless Communications and Mobile Computing* 2018 (Aug. 2018), pp. 1–23, ISSN: 1530-8669, 1530-8677, DOI: 10.1155/2018/8035065, URL: https://www.hindawi.com/journals/wcmc/2018/8035065/ (visited on 11/30/2022).

[ES07] Adir Even and G. Shankaranarayanan, "Utility-driven assessment of data quality", en, *in*: *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 38.*2* (May 2007), pp. 75–93, ISSN: 0095-0033, 1532-0936, DOI: 10.1145/1240616.1240623, URL: https://dl.acm.org/doi/10.1145/1240616.1240623 (visited on 06/08/2023).

[Fas+07] Elena Fasolo et al., "In-network aggregation techniques for wireless sensor networks: a survey", en, *in*: *IEEE Wireless Communications* 14.*2* (Apr. 2007), pp. 70–87, ISSN: 1536-1284, DOI: 10.1109/MWC.2007.358967, URL: http://ieeexplore.ieee.org/document/4198169/ (visited on 12/01/2022).

[Gez08] Sinan Gezici, "A Survey on Wireless Position Estimation", en, *in*: *Wireless Personal Communications* 44.*3* (Feb. 2008), pp. 263–282, ISSN: 1572-834X, DOI: 10.1007/s11277-007-9375-z, URL: https://doi.org/10.1007/s11277-007-9375-z (visited on 05/11/2023).

[GGJ09] H. Ghasemzadeh, E. Guenterberg, and R. Jafari, "Energy-Efficient Information-Driven Coverage for Physical Movement Monitoring in Body Sensor Networks", en, *in*: *IEEE Journal on Selected Areas in Communications* 27.*1* (Jan. 2009), pp. 58–69, ISSN: 0733-8716, DOI: 10.1109/JSAC.2009.090107, URL: http://ieeexplore.ieee.org/document/4740886/ (visited on 11/29/2022).

[GR17] Anna Großwendt and Heiko Röglin, "Improved Analysis of Complete-Linkage Clustering", en, *in*: *Algorithmica* 78.*4* (Aug. 2017), pp. 1131–1150, ISSN: 1432-0541, DOI: 10.1007/s00453-017-0284-6, URL: https://doi.org/10.1007/s00453-017-0284-6 (visited on 07/17/2023).

[Gru+06] Jaap J. de Gruijter et al., *Sampling for Natural Resource Monitoring*, en, Berlin, Heidelberg: Springer, 2006, ISBN: 978-3-540-22486-0 978-3-540-33161-2, DOI: 10.1007/3-540-33161-1, URL: http://link.springer.com/10.1007/3-540-33161-1 (visited on 04/14/2023).

[GTG22] Srishti Gupta, Sarvesh Tanwar, and Neelam Gupta, "A Systematic Review on Internet of Things (IoT): Applications & Challenges", *in*: *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Oct. 2022, pp. 1–7, DOI: 10.1109/ICRITO56286.2022.9964892.

[Gue11] Pamela Guevara, "Inference of a human brain fiber bundle atlas from high angular resolution diffusion imaging", *in*: (Oct. 2011).

[GZD21] Panagiotis Gkotsiopoulos, Dimitrios Zorbas, and Christos Douligeris, "Performance Determinants in LoRa Networks: A Literature Review", *in*: *IEEE Communications Surveys & Tutorials* 23.*3* (2021), Conference Name: IEEE Communications Surveys & Tutorials, pp. 1721–1758, ISSN: 1553-877X, DOI: 10.1109/COMST.2021.3090409.

[HLR09] Michael A. Henning, Christian Löwenstein, and Dieter Rautenbach, "Remarks about disjoint dominating sets", en, *in*: *Discrete Mathematics* 309.*23* (Dec. 2009), pp. 6451–6458, ISSN: 0012-365X, DOI: 10.1016/j.disc.2009.06.017, URL: https://www.sciencedirect.com/science/article/pii/S0012365X09003288 (visited on 02/23/2023).

[JC19] Aarti Jangid and Parul Chauhan, "A Survey and Challenges in IoT Networks", *in*: *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, Feb. 2019, pp. 516–521, DOI: 10.1109/ISS1.2019.8908079.

[Jou+23] Mohammed Jouhari et al., "A Survey on Scalable LoRaWAN for Massive IoT: Recent Advances, Potentials, and Challenges", *in*: *IEEE Communications Surveys & Tutorials* (2023), Conference Name: IEEE Communications Surveys & Tutorials, pp. 1–1, ISSN: 1553-877X, DOI: 10.1109/COMST.2023.3274934.

[JR22] Akram H. Jebril and Rozeha A. Rashid, "A systematic literature review on downlink frames in LoRaWAN", en, *in*: *Computers and Electrical Engineering* 101 (July 2022), p. 108006, ISSN: 0045-7906, DOI: 10.1016/j.

compeleceng.2022.108006, URL: https://www.sciencedirect.com/science/article/pii/S0045790622002737 (visited on 05/11/2023).

[Kai+16] Omprakash Kaiwartya et al., "T-MQM: Testbed-Based Multi-Metric Quality Measurement of Sensor Deployment for Precision Agriculture—A Case Study", *in*: *IEEE Sensors Journal* 16.*23* (Dec. 2016), Conference Name: IEEE Sensors Journal, pp. 8649–8664, ISSN: 1558-1748, DOI: 10.1109/JSEN.2016.2614748.

[Kar+16] Aimad Karkouch et al., "Data quality in internet of things: A state-of-the-art survey", en, *in*: *Journal of Network and Computer Applications* 73 (Sept. 2016), pp. 57–81, ISSN: 10848045, DOI: 10.1016/j.jnca.2016.08.002, URL: https://linkinghub.elsevier.com/retrieve/pii/S1084804516301564 (visited on 11/28/2022).

[KAT12] Vibhav KumarSachan, Syed Akhtar Imam, and M. T. Beg, "Energy-Efficient Communication Methods in Wireless Sensor Networks: A Critical Review", en, *in*: *International Journal of Computer Applications* 39.*17* (Feb. 2012), pp. 35–48, ISSN: 09758887, DOI: 10.5120/4915-7484, URL: http://research.ijcaonline.org/volume39/number17/pxc3877484.pdf (visited on 11/29/2022).

[KGG11] Petr Kadlec, Ratko Grbić, and Bogdan Gabrys, "Review of adaptation mechanisms for data-driven soft sensors", en, *in*: *Computers & Chemical Engineering* 35.*1* (Jan. 2011), pp. 1–24, ISSN: 0098-1354, DOI: 10.1016/j.compchemeng.2010.07.034, URL: https://www.sciencedirect.com/science/article/pii/S0098135410002838 (visited on 04/26/2023).

[KL05] N. Kimura and S. Latifi, "A survey on data compression in wireless sensor networks", en, *in*: *International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II*, Las Vegas, NV, USA: IEEE, 2005, 8–13 Vol. 2, ISBN: 978-0-7695-2315-6, DOI: 10.1109/ITCC.2005.43, URL: http://ieeexplore.ieee.org/document/1425113/ (visited on 12/01/2022).

[Kle09] Jack P.C. Kleijnen, "Kriging metamodeling in simulation: A review", en, *in*: *European Journal of Operational Research* 192.*3* (Feb. 2009), pp. 707–716, ISSN: 03772217, DOI: 10.1016/j.ejor.2007.10.013, URL: https://linkinghub.elsevier.com/retrieve/pii/S0377221707010090 (visited on 07/11/2023).

[Kno06] R. Knopp, "Two-Way Radio Networks with a Star Topology", *in*: *2006 International Zurich Seminar on Communications*, Feb. 2006, pp. 154–157, DOI: 10.1109/IZS.2006.1649103.

[KPK97] R.M. Kil, Seon Hee Park, and Seunghwan Kim, "Optimum window size for time series prediction", *in*: *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 'Magnificent Milestones and Emerging Opportunities in Medical Engineering' (Cat. No.97CH36136)*, vol. 4, ISSN: 1094-687X, Oct. 1997, 1421–1424 vol.4, DOI: 10.1109/IEMBS.1997.756971.

[Kra+06] A. Krause et al., "Near-optimal sensor placements: maximizing information while minimizing communication cost", *in*: *2006 5th International Conference on Information Processing in Sensor Networks*, Apr. 2006, pp. 2–10, DOI: 10.1145/1127777.1127782.

[Kra+11] Andreas Krause et al., "Simultaneous Optimization of Sensor Placements and Balanced Schedules", en, *in*: *IEEE Transactions on Automatic Control* 56.*10* (Oct. 2011), pp. 2390–2405, ISSN: 0018-9286, 1558-2523, DOI: 10.1109/TAC.2011.2164010, URL: http://ieeexplore.ieee.org/document/5978192/ (visited on 11/29/2022).

[KS17] Navroop Kaur and Sandeep K. Sood, "An Energy-Efficient Architecture for the Internet of Things (IoT)", en, *in*: *IEEE Systems Journal* 11.*2* (June 2017), pp. 796–805, ISSN: 1932-8184, 1937-9234, 2373-7816, DOI: 10.1109/JSYST.2015.2469676, URL: http://ieeexplore.ieee.org/document/7293596/ (visited on 11/28/2022).

[KTP06] F. Koushanfar, N. Taft, and M. Potkonjak, "Sleeping Coordination for Comprehensive Sensing Using Isotonic Regression and Domatic Partitions", en, *in*: *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, Barcelona, Spain: IEEE, 2006, pp. 1–13, ISBN: 978-1-4244-0221-2, DOI: 10.1109/INFOCOM.2006.276, URL: http://ieeexplore.ieee.org/document/4146929/ (visited on 11/28/2022).

[Kum+17] Kirshna Kumar et al., "Cross-Layer Energy Optimization for IoT Environments: Technical Advances and Opportunities", en, *in*: *Energies* 10.*12* (Dec. 2017), p. 2073, ISSN: 1996-1073, DOI: 10.3390/en10122073, URL: http://www.mdpi.com/1996-1073/10/12/2073 (visited on 11/28/2022).

[Lal+18] Guénolé Lallement et al., "A 2.7 pJ/cycle 16 MHz, 0.7 mu W Deep Sleep Power ARM Cortex-M0+ Core SoC in 28 nm FD-SOI", *in*: *IEEE Journal of Solid-State Circuits* 53.*7* (July 2018), Conference Name: IEEE Journal of Solid-State Circuits, pp. 2088–2100, ISSN: 1558-173X, DOI: 10.1109/JSSC.2018.2821167.

[Lia05] T. Warren Liao, "Clustering of time series data-a survey", en, *in*: *Pattern Recognition* (Nov. 2005), DOI: 10.1016/j.patcog.2005.01.025, URL: https://www.scinapse.io/papers/2097747115 (visited on 07/18/2023).

[Liu+13a] Zhidan Liu et al., "An energy-efficient data collection scheme for wireless sensor networks", *in*: *2013 15th International Conference on Advanced Communications Technology (ICACT)*, ISSN: 1738-9445, Jan. 2013, pp. 60–65.

[Liu+13b] Zhidan Liu et al., "Hierarchical Spatial Clustering in Multihop Wireless Sensor Networks", *in*: *International Journal of Distributed Sensor Networks* 9.*11* (Nov. 2013), Publisher: SAGE Publications, p. 528980, ISSN: 1550-1329, DOI: 10.1155/2013/528980, URL: https://doi.org/10.1155/2013/528980 (visited on 11/30/2022).

[LW67] G. N. Lance and W. T. Williams, "A general theory of classificatory sorting strategies: II. Clustering systems", *in*: *The Computer Journal* 10.*3* (Jan. 1967), pp. 271–277, ISSN: 0010-4620, DOI: 10.1093/comjnl/10.3.271, URL: https://doi.org/10.1093/comjnl/10.3.271 (visited on 07/18/2023).

[LWP07] Chong Liu, Kui Wu, and Jian Pei, "An Energy-Efficient Data Collection Framework for Wireless Sensor Networks by Exploiting Spatiotemporal Correlation", *in*: *IEEE Transactions on Parallel and Distributed Systems* 18.*7* (July 2007), Conference Name: IEEE Transactions on Parallel and Distributed Systems, pp. 1010–1023, ISSN: 1558-2183, DOI: 10.1109/TPDS.2007.1046.

[Mam14] Quazi Mamun, "A Coverage-Based Scheduling Algorithm for WSNs", en, *in*: *International Journal of Wireless Information Networks* 21.*1* (Mar. 2014), pp. 48–57, ISSN: 1068-9605, 1572-8129, DOI: 10.1007/s10776-013-0231-7, URL: http://link.springer.com/10.1007/s10776-013-0231-7 (visited on 12/06/2022).

[Mau+22] Gwen Maudet et al., "Emission Scheduling Strategies for Massive-IoT: Implementation and Performance Optimization", *in*: *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*, ISSN: 2374-9709, Apr. 2022, pp. 1–4, DOI: 10.1109/NOMS54207.2022.9789769.

[Mau+23] Gwen Maudet et al., "Energy Efficient Message Scheduling with Redundancy Control for Massive IoT Monitoring", en, *in*: *2023 IEEE Wireless Communications and Networking, 2023. WCNC 2023.* 2023.

[MC12] Fionn Murtagh and Pedro Contreras, "Algorithms for hierarchical clustering: an overview", en, *in*: *WIREs Data Mining and Knowledge Discovery* 2.*1* (2012), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.53, pp. 86–97, ISSN: 1942-4795, DOI: 10.1002/widm.53, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.53 (visited on 07/19/2023).

[Meg+01] S. Meguerdichian et al., "Coverage problems in wireless ad-hoc sensor networks", *in*: *Proceedings IEEE INFOCOM 2001. Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society (Cat. No.01CH37213)*, vol. 3, ISSN: 0743-166X, Apr. 2001, 1380–1387 vol.3, DOI: 10.1109/INFCOM.2001.916633.

[Mek+19] Kais Mekki et al., "A comparative study of LPWAN technologies for large-scale IoT deployment", en, *in*: *ICT Express* 5.*1* (Mar. 2019), pp. 1–7, ISSN: 24059595, DOI: 10.1016/j.icte.2017.12.005, URL: https://linkinghub.elsevier.com/retrieve/pii/S2405959517302953 (visited on 12/02/2022).

[Mey+11] J.m. Meyers et al., "Improving vineyard sampling efficiency via dynamic spatially explicit optimisation", en, *in*: *Australian Journal of Grape and Wine Research* 17.*3* (2011), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-0238.2011.00152.x, pp. 306–315, ISSN: 1755-0238, DOI: 10.1111/j.1755-0238.2011.00152.x, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0238.2011.00152.x (visited on 05/11/2023).

[MF09] Guoqiang Mao and Baris Fidan, "Localization Algorithms and Strategies for Wireless Sensor Networks", en, *in*: *Information Science Reference* (2009), DOI: 10.4018/978-1-60566-396-8.

[MM08] Saoucene Mahfoudh and Pascale Minet, "Survey of Energy Efficient Strategies in Wireless Ad Hoc and Sensor Networks", en, *in*: *Seventh International Conference on Networking (icn 2008)*, Cancun, Mexico: IEEE, Apr. 2008, pp. 1–7, ISBN: 978-0-7695-3106-9, DOI: 10.1109/ICN.2008.55, URL: http://ieeexplore.ieee.org/document/4498134/ (visited on 12/01/2022).

[MMR05] Oded Maimon, Oded Z. Maimon, and Lior Rokach, *Data Mining and Knowledge Discovery Handbook*, en, Google-Books-ID: jizrAIWUJ6UC, Springer Science & Business Media, 2005, ISBN: 978-0-387-24435-8.

[Mos+17] Habib Mostafaei et al., "A sleep scheduling approach based on learning automata for WSN partialcoverage", en, *in*: *Journal of Network and Computer Applications* 80 (Feb. 2017), pp. 67–78, ISSN: 1084-8045, DOI: 10.1016/j.jnca.2016.12.022, URL: https://www.sciencedirect.com/science/article/pii/S1084804516303204 (visited on 05/04/2023).

[Mot+20] Naser Hossein Motlagh et al., "Toward Massive Scale Air Quality Monitoring", *in*: *IEEE Communications Magazine* 58.*2* (Feb. 2020), Conference Name: IEEE Communications Magazine, pp. 54–59, ISSN: 1558-1896, DOI: 10.1109/MCOM.001.1900515.

[NC23] Ihab Nassra and Juan V. Capella, "Data Compression Techniques in IoT-enabled Wireless Body Sensor Networks: A Systematic Literature Review and Research Trends for QoS Improvement", en, *in*: *Internet of Things* (May 2023), p. 100806, ISSN: 2542-6605, DOI: 10.1016/j.iot.2023.100806, URL: https://www.sciencedirect.com/science/article/pii/S2542660523001294 (visited on 05/09/2023).

[NH07] Trac N. Nguyen and Dung T. Huynh, "Extending Sensor Networks Lifetime Through Energy Efficient Organization", en, *in*: *International Conference on Wireless Algorithms, Systems and Applications (WASA 2007)*, Chicago, IL: IEEE, Aug. 2007, pp. 205–212, ISBN: 978-0-7695-2981-3, DOI: 10.1109/WASA.2007.7, URL: http://ieeexplore.ieee.org/document/4288232/ (visited on 11/29/2022).

[NWN15] Hai-Long Nguyen, Yew-Kwong Woon, and Wee-Keong Ng, "A survey on data stream clustering and classification", en, *in*: *Knowledge and Information Systems* 45.*3* (Dec. 2015), pp. 535–569, ISSN: 0219-1377, 0219-3116, DOI:

10.1007/s10115-014-0808-1, URL: http://link.springer.com/10.1007/s10115-014-0808-1 (visited on 12/01/2022).

[OQ09] Olawoye Oyeyele and Hairong Qi, "A Robust Node Selection Strategy for Lifetime Extension in Wireless Sensor Networks", *in*: *2009 Fifth International Conference on Mobile Ad-hoc and Sensor Networks*, Dec. 2009, pp. 196–203, DOI: 10.1109/MSN.2009.84.

[Per+20a] Felisberto Pereira et al., "Challenges in Resource-Constrained IoT Devices: Energy and Communication as Critical Success Factors for Future IoT Deployment", en, *in*: *Sensors* 20.*22* (Jan. 2020), Number: 22 Publisher: Multidisciplinary Digital Publishing Institute, p. 6420, ISSN: 1424-8220, DOI: 10.3390/s20226420, URL: https://www.mdpi.com/1424-8220/20/22/6420 (visited on 05/11/2023).

[Per+20b] Toni Perković et al., "Meeting Challenges in IoT: Sensing, Energy Efficiency, and the Implementation", en, *in*: *Fourth International Congress on Information and Communication Technology*, ed. by Xin-She Yang et al., Advances in Intelligent Systems and Computing, Singapore: Springer, 2020, pp. 419–430, ISBN: 9789811506376, DOI: 10.1007/978-981-15-0637-6_36.

[PSP22] Amitkumar Patil, Gunjan Soni, and Anuj Prakash, "Data-driven approaches for impending fault detection of industrial systems: a review", en, *in*: *International Journal of System Assurance Engineering and Management* (Dec. 2022), ISSN: 0976-4348, DOI: 10.1007/s13198-022-01841-9, URL: https://doi.org/10.1007/s13198-022-01841-9 (visited on 04/26/2023).

[PT00] S. Punronen and V. Terziyan, "A similarity evaluation technique for data mining with an ensemble of classifiers", *in*: *Proceedings 11th International Workshop on Database and Expert Systems Applications*, ISSN: 1529-4188, Sept. 2000, pp. 1155–1159, DOI: 10.1109/DEXA.2000.875172.

[PWG13] Victor Picheny, Tobias Wagner, and David Ginsbourger, "A benchmark of kriging-based infill criteria for noisy optimization", en, *in*: *Structural and Multidisciplinary Optimization* 48.*3* (Sept. 2013), pp. 607–626, ISSN: 1615-147X, 1615-1488, DOI: 10.1007/s00158-013-0919-4, URL: http://link.springer.com/10.1007/s00158-013-0919-4 (visited on 07/11/2023).

[Raj11] Ramesh Rajagopalan, "Spatial Correlation Based Sensor Selection Schemes for Probabilistic Area Coverage", en, *in*: *International journal of Computer Networks & Communications* 3.*2* (Mar. 2011), pp. 233–249, ISSN: 09752293, DOI: 10.5121/ijcnc.2011.3215, URL: http://www.airccse.org/journal/cnc/0311cnc15.pdf (visited on 11/29/2022).

[Ran71] William M. Rand, "Objective Criteria for the Evaluation of Clustering Methods", *in*: *Journal of the American Statistical Association* 66.*336* (1971), Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 846–850, ISSN: 0162-1459, DOI: 10.2307/2284239, URL: https://www.jstor.org/stable/2284239 (visited on 07/18/2023).

[RBC14] Tifenn Rault, Abdelmadjid Bouabdallah, and Yacine Challal, "Energy efficiency in wireless sensor networks: A top-down survey", en, *in*: *Computer Networks* 67 (July 2014), pp. 104–122, ISSN: 13891286, DOI: 10.1016/j.comnet.2014.03.027, URL: https://linkinghub.elsevier.com/retrieve/pii/S1389128614001418 (visited on 11/28/2022).

[RC23] Husam Rajab and Tibor Cinkler, "Enhanced Energy Efficiency and Scalability in Cellular Networks for Massive IoT", en, *in*: *5G and Beyond*, ed. by Bharat Bhushan et al., Springer Tracts in Electrical and Electronics Engineering, Singapore: Springer Nature, 2023, pp. 283–305, ISBN: 978-981-9936-68-7, DOI: 10.1007/978-981-99-3668-7_13, URL: https://doi.org/10.1007/978-981-99-3668-7_13 (visited on 09/04/2023).

[SC19] Yin Sun and Benjamin Cyr, "Sampling for data freshness optimization: Nonlinear age functions", *in*: *Journal of Communications and Networks* 21.*3* (June 2019), Conference Name: Journal of Communications and Networks, pp. 204–219, ISSN: 1976-5541, DOI: 10.1109/JCN.2019.000035.

[Shi+01] Eugene Shih et al., "Physical layer driven protocol and algorithm design for energy-efficient wireless sensor networks", en, *in*: *Proceedings of the 7th annual international conference on Mobile computing and networking*, Rome Italy: ACM, July 2001, pp. 272–287, ISBN: 978-1-58113-422-3, DOI: 10.1145/381677.381703, URL: https://dl.acm.org/doi/10.1145/381677.381703 (visited on 05/16/2023).

[SHS01] Andreas Savvides, Chih-Chieh Han, and Mani B. Strivastava, "Dynamic fine-grained localization in Ad-Hoc networks of sensors", *in*: *Proceedings of the 7th annual international conference on Mobile computing and networking*, MobiCom '01, New York, NY, USA: Association for Computing Machinery, 2001, pp. 166–179, ISBN: 978-1-58113-422-3, DOI: 10.1145/381677.381693, URL: https://doi.org/10.1145/381677.381693 (visited on 01/03/2023).

[SP01] S. Slijepcevic and M. Potkonjak, "Power efficient organization of wireless sensor networks", *in*: *ICC 2001. IEEE International Conference on Communications. Conference Record (Cat. No.01CH37240)*, vol. 2, June 2001, 472–476 vol.2, DOI: 10.1109/ICC.2001.936985.

[SSV12] Rajeev K. Shakya, Yatindra Nath Singh, and Nishchal K. Verma, "A novel spatial correlation model for wireless sensor network applications", en, *in*: *2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)*, Indore, India: IEEE, Sept. 2012, pp. 1–6, ISBN: 978-1-4673-1989-8 978-1-4673-1988-1 978-1-4673-1987-4, DOI: 10.1109/WOCN.2012.6335549, URL: http://ieeexplore.ieee.org/document/6335549/ (visited on 12/01/2022).

[Stu+19] Martin Stusek et al., "IoT Protocols for Low-power Massive IoT: A Communication Perspective", *in*: *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, ISSN: 2157-023X, Oct. 2019, pp. 1–7, DOI: 10.1109/ICUMT48472.2019.8970868.

[Tao+23] Zhifu Tao et al., "An integrated approach implementing sliding window and DTW distance for time series forecasting tasks", en, *in*: *Applied Intelligence* (Apr. 2023), ISSN: 1573-7497, DOI: 10.1007/s10489-023-04590-9, URL: https://doi.org/10.1007/s10489-023-04590-9 (visited on 04/24/2023).

[TC07] Maen Takruri and Subhash Challa, "Drift aware wireless sensor networks", *in*: *2007 10th International Conference on Information Fusion*, July 2007, pp. 1–7, DOI: 10.1109/ICIF.2007.4408091.

[TDT07] Andreas Tolk, Saikou Diallo, and Charles Turnitsa, "Applying the Levels of Conceptual Interoperability Model in Support of Integratability, Interoperability, and Composability for System-of-Systems Engineering", *in*: *International Journal Systemics, Cybernetics and Informatics* 5 (Oct. 2007).

[TG02] Di Tian and Nicolas D Georganas, "A Coverage-Preserving Node Scheduling Scheme for Large Wireless Sensor Networks", en, *in*: *Proceedings of the 1st ACM International Wrkshopt on Wireless Sensor Networks and Applications* (2002), p. 10, DOI: 10.1145/570738.570744.

[TG03] Di Tian and Nicolas D. Georganas, "A node scheduling scheme for energy conservation in large wireless sensor networks", en, *in*: *Wireless Communications and Mobile Computing* 3.*2* (2003), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1 pp. 271–290, ISSN: 1530-8677, DOI: 10.1002/wcm.116, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcm.116 (visited on 01/03/2023).

[THH02] Yu-Chee Tseng, Chih-Shun Hsu, and Ten-Yueng Hsieh, "Power-saving protocols for IEEE 802.11-based multi-hop ad hoc networks", *in*: *Proceedings.Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 1, ISSN: 0743-166X, June 2002, 200–209 vol.1, DOI: 10.1109/INFCOM.2002.1019261.

[Tjo+04] R. Tjoa et al., "Clock drift reduction for relative time slot TDMA-based sensor networks", *in*: *2004 IEEE 15th International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE Cat. No.04TH8754)*, vol. 2, Sept. 2004, 1042–1047 Vol.2, DOI: 10.1109/PIMRC.2004.1373857.

[TM06] Daniela Tulone and Samuel Madden, "An energy-efficient querying framework in sensor networks for detecting node similarities", en, *in*: *Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems - MSWiM '06*, Terromolinos, Spain: ACM Press, 2006, p. 191, ISBN: 978-1-59593-477-2, DOI: 10.1145/1164717.1164768, URL: http://portal.acm.org/citation.cfm?doid=1164717.1164768 (visited on 11/30/2022).

[Tol03] Andreas Tolk, "Levels of Conceptual Interoperability", en, *in*: (2003), p. 11.

[US20] Silvia Liberata Ullo and G. R. Sinha, "Advances in Smart Environment Monitoring Systems Using IoT and Sensors", en, *in*: *Sensors* 20.*11* (May 2020), p. 3113, ISSN: 1424-8220, DOI: 10.3390/s20113113, URL: https://www.mdpi.com/1424-8220/20/11/3113 (visited on 12/02/2022).

[VAA04] Mehmet C. Vuran, Ozgur B. Akan, and Ian F. Akyildiz, "Spatio-temporal correlation: theory and applications for wireless sensor networks", en, *in*: *Computer Networks* 45.*3* (June 2004), pp. 245–259, ISSN: 13891286, DOI: 10.1016/j.comnet.2004.03.007, URL: https://linkinghub.elsevier.com/retrieve/pii/S1389128604000519 (visited on 12/01/2022).

[Vaf+14] Majid Vafaeipour et al., "Application of sliding window technique for prediction of wind velocity time series", en, *in*: *International Journal of Energy and Environmental Engineering* 5.*2* (May 2014), p. 105, ISSN: 2251-6832, DOI: 10.1007/s40095-014-0105-5, URL: https://doi.org/10.1007/s40095-014-0105-5 (visited on 04/24/2023).

[Wan+13] Jin-Feng Wang et al., "Design-based spatial sampling: Theory and implementation", en, *in*: *Environmental Modelling & Software* 40 (Feb. 2013), pp. 280–288, ISSN: 1364-8152, DOI: 10.1016/j.envsoft.2012.09.015, URL: https://www.sciencedirect.com/science/article/pii/S1364815212002502 (visited on 05/11/2023).

[WX06] Lan Wang and Yang Xiao, "A Survey of Energy-Efficient Scheduling Mechanisms in Sensor Networks", en, *in*: *Mobile Networks and Applications* 11.*5* (Oct. 2006), pp. 723–740, ISSN: 1383-469X, 1572-8153, DOI: 10.1007/s11036-006-7798-5, URL: http://link.springer.com/10.1007/s11036-006-7798-5 (visited on 11/29/2022).

[YA08] Mohamed Younis and Kemal Akkaya, "Strategies and techniques for node placement in wireless sensor networks: A survey", en, *in*: *Ad Hoc Networks* 6.*4* (June 2008), pp. 621–655, ISSN: 15708705, DOI: 10.1016/j.adhoc.2007.05.003, URL: https://linkinghub.elsevier.com/retrieve/pii/S1570870507000984 (visited on 11/29/2022).

[Yog+18] Duwaraka Yoganathan et al., "Optimal sensor placement strategy for office buildings using clustering algorithms", en, *in*: *Energy and Buildings* 158 (Jan. 2018), pp. 1206–1225, ISSN: 0378-7788, DOI: 10.1016/j.enbuild.2017.10.074, URL: https://www.sciencedirect.com/science/article/pii/S0378778817304723 (visited on 04/26/2023).

[Yoo20] YoungJun Yoo, "Data-driven fault detection process using correlation based clustering", *in*: *Computers in Industry* 122 (Nov. 2020), MAG ID: 3040582905

S2ID: ffb75bed2efc2b4d823a209272f40637145d7ab9, p. 103279, DOI: 10.1016/j.compind.2020.103279.

[Zan+21] Eljona Zanaj et al., "Energy Efficiency in Short and Wide-Area IoT Technologies—A Survey", en, *in*: *Technologies* 9.*1* (Mar. 2021), p. 22, ISSN: 2227-7080, DOI: 10.3390/technologies9010022, URL: https://www.mdpi.com/2227-7080/9/1/22 (visited on 11/30/2022).

[Zhe+20] Xiulin Zheng et al., "A Survey on Multi-Label Data Stream Classification", en, *in*: *IEEE Access* 8 (2020), pp. 1249–1275, ISSN: 2169-3536, DOI: 10.1109/ACCESS.2019.2962059, URL: https://ieeexplore.ieee.org/document/8941052/ (visited on 12/01/2022).

[Zhu+12] Chuan Zhu et al., "A survey on coverage and connectivity issues in wireless sensor networks", en, *in*: *Journal of Network and Computer Applications* 35.*2* (Mar. 2012), pp. 619–632, ISSN: 10848045, DOI: 10.1016/j.jnca.2011.11.016, URL: https://linkinghub.elsevier.com/retrieve/pii/S1084804511002323 (visited on 11/29/2022).

[Zim+99] Dale Zimmerman et al., "An Experimental Comparison of Ordinary and Universal Kriging and Inverse Distance Weighting", en, *in*: *Mathematical Geology* 31.*4* (1999), pp. 375–390, ISSN: 08828121, DOI: 10.1023/A:1007586507433, URL: http://link.springer.com/10.1023/A:1007586507433 (visited on 07/11/2023).

COLLEGES BRETAGNE LOIRE
SCIENCES POUR L'INGENIEUR ET LE NUMERIQUE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Exploiter la Similarité des Capteurs pour Améliorer la Collecte de Données dans les Réseaux IoT Massifs

Mot clés : Internet des Objets Massif, Efficacité pour l'IoT, Gestion des Collections d'Observations

Résumé : L'Internet des objets (IoT) est couramment utilisé pour surveiller diverses grandeurs physiques. Dans l'approche innovante du Massive IoT (MIoT), un déploiement massif de capteurs très contraints est envisagé, afin de réduire les coûts de déploiement et de maintenance. Conformément à ce scénario, cette thèse se concentre sur le développement de mécanismes visant à réduire la consommation d'énergie des capteurs. La méthode repose sur le principe de similarité : les capteurs peuvent être considérés comme similaires s'ils fournissent des observations semblables. Cette approche permet la transmission d'un sous-ensemble de capteurs répondant aux exigences de surveillance. Tout d'abord, nous avons identifié et synthétisé les méthodes existantes provenant de la littérature basées sur le principe de similarité. Nous avons etablit que ce type d'approche peut être décomposé en trois composantes, que nous avons étudiées dans le contexte du MIoT. Ensuite, nous avons examiné les méthodes de gestion des observations des capteurs permettant de maintenir une quantité constante de messages au fil du temps. Notre première méthode permet de transmettre en round-robin un nombre spécifié de capteurs. La deuxième méthode atteint des résultats de précision comparables à la première tout en réduisant le nombre de mises à jour des capteurs lorsque la flotte de capteurs change. Enfin, nous proposons une solution pour former des groupes de capteurs identifiés comme similaires en analysant leurs observations. À cet effet, nous introduisons une nouvelle mesure de similarité basée sur l'interpolation, associée à une méthode de regroupement hiérarchique.

Title: Exploiting Sensor Similarity to Enhance Data Collection in Massive IoT Networks

Keywords: Massive Internet of Thing, IoT Efficiency, Observation Collection Management

Abstract: The Internet of Things (IoT) are commonly employed for monitoring various physical quantities. In the innovative approach of Massive IoT (MIoT), a massive deployment of highly constrained sensors is considered to reduce deployment and maintenance costs. Aligned with this scenario, this thesis focuses on the development of mechanisms to reduce sensor energy consumption. The method relies on the principle of similarity: sensors can be considered similar if they provide similar observations. This approach enables the transmission of a subset of sensors to fulfill the monitoring requirements. First, we identified and synthesized existing methods from the literature based on the principle of similarity. We established that this approach can be decomposed into three components, which we studied in the context of MIoT. Next, we examined methods for managing sensor observations to maintain a constant stream of messages over time. Our first method involves transmitting a specified number of sensors in a round-robin fashion. The second method achieves precision results comparable to the first while reducing the number of sensor updates when the sensor fleet changes. Finally, we propose a solution to form groups of sensors identified as similar by analyzing their observations. To this end, we introduce a new similarity measure based on interpolation, coupled with a hierarchical clustering method.